Perceptual Categorization

Neil Stewart

Thesis submitted for the degree of Doctor of Philosophy

Department of Psychology
University of Warwick

June 2001

Contents

Table of Contents

List of Figures

List of Tables

Acknowledgements

I would like to thank Nick Chater and Evan Heit for their supportive supervision of this thesis. Their advice and comments were invaluable. I am also grateful to the other members of the Psychology Department at Warwick for their help. Particular thanks are to rightfully be given to Gordon Brown, with whom I have spent many hours discussing the ideas presented here. Thanks also go to Suzanna Bootle, Lewis Bott, Jon Brock, Eoghan Clarkson, Koen Lamberts and Derrick Watson. Philippe Schyns and Annie Archambault of the University of Glasgow also deserve thanks for supplying the stimuli used in Experiment 9, and providing helpful comments on how to make the experiment work! Discussion with Stian Reimers of the University of Cambridge at the Star Burger conferences held in Coventry and Cambridge has played a crucial role supporting the development of ideas in this thesis.

I was supported financially by a University of Warwick Graduate Assistantship. Additional funds for testing participants came from a BBSRC grant held by Evan Heit, and from the Department of Psychology.

Finally, I am very grateful to Catherine, for putting up with me.

Declaration

I hereby declare that the research reported in this thesis is my own work unless otherwise stated. No part of this thesis has been submitted for a degree at another university.

Chapter 2 was written in collaboration with Nick Chater. Chapter 3 was written in collaboration with Gordon Brown and Nick Chater. The "memory and contrast" model described in Chapter 3 was developed by myself and Gordon Brown. Two experiments in Chapter 3 (Experiments 6 and 8) were run by Suzanna Bootle. Both chapters have been submitted to the Journal of Experimental Psychology: Learning, Memory and Cognition. Suggestions for further work in Chapter 5 are based on two grant proposals written in collaboration with Gordon Brown, Nick Chater and Koen Lamberts, and Derrick Watson.

Neil Stewart

Abstract

The categorization of external stimuli lies at the heart of cognitive science. Existing models of perceptual categorization assume (a) information about the absolute magnitude of a stimulus is used in the categorization decision, and (b) the representation of a stimulus does not change with experience. The three experimental programs presented here challenge these two assumptions. The experiments in Chapter 2 demonstrate that existing models of categorization are unable to predict the classification of items intermediate between two categories. Chapter 3 provides empirical evidence that categorization responses are heavily influenced by the immediately preceding context, consistent with evidence from absolute identification showing people have very poor access to absolute magnitude information. A memory and contrast model is presented where each categorization decision is based on the perceived difference between the current stimulus and immediately preceding stimuli. This model is shown to account for the data from Chapters 2 and 3. Chapter 4 explores the claim that new features may be created on experience with novel stimuli, and that these features serve to alter the representation of stimuli to facilitate new categorization tasks. An alternative account is offered for existing feature creation evidence. However, experimental work re-establishes a feature creation effect. Consideration is given as to how feature creation and memory and contrast accounts of categorization may be integrated, together with extensive suggestions for the development of these ideas.

Chapter 1

Introduction

Perceptual categorization involves the grouping of individual, discriminable items together. Novel items may be judged members of these categories, and properties of category members generalized to these novel items. A distinction may be drawn between how categories are represented, and how the representations are used to classify novel items. The distinction then is between information availability and information use. Theories of categorization have been instantiated as mathematical models of categorization that make explicit the assumptions about both representation and process. However, much of the literature focuses on the nature of the representation. This issue lies at the heart of cognitive science, for it is central to the link between perception and cognition. The development of theories of categorization is therefore largely concerned with the representation of categories. This chapter provides a review of existing models of perceptual categorization, and related experimental evidence.

The classical theory of categorization (so named by Bruner, Goodnow, & Austin, 1956; Smith & Medin, 1981) states that items are categorized into groups on the basis of a list of necessary and sufficient features or attributes. If an item possesses all the attributes in the category's list then it is said to be a category member. Wittgenstein (1958) pointed out that for many natural categories exemplars do not all share the same common attribute or property. Further, if necessary and sufficient features are to be the basis for categorization, the classical viewpoint suggests that all exemplars of a category should be equally good examples of that category.

However, not all exemplars are judged to be equally good examples of their category. Rosch (1976) found that people were faster to verify category membership statements involving more typical members of a natural category. For example,

people are faster to verify "a canary is a bird" than "an emu is a bird". This typicality effect is also true of categories of novel or artificial stimuli. For example, unseen prototypes can be categorized more accurately than the original training stimuli (Estes, 1986; Hintzman, 1986; Homa, Sterling, & Trepel, 1981; Lamberts, 1996; Medin & Schaffer, 1978). Category membership may then be better considered in terms of family resemblance (Rosch & Mervis, 1975). This idea prompted the development of prototype models of categorization (e.g., Homa et al., 1981; Posner & Keele, 1968; Posner & Keele, 1970; Reed, 1972; Rosch, 1973; Rosch et al., 1976). In prototype theories category membership is a matter of degree, and is based on an exemplar's similarity to the average category member or central tendency of the category.

The prototype account does not rest easily with other empirical findings. A large number of experiments have demonstrated that performance in categorization tasks can be influenced by exemplars other than the category prototype (e.g., Ashby & Gott, 1988; Brooks, 1978; Medin & Schaffer, 1978; Whittlesea, 1987). Old training stimuli can be categorized faster than new, previously unseen stimuli equally similar to the prototype (Jacoby & Brooks, 1984). Malt (1989) showed that the categorization of an old exemplar could be primed by prior presentation of a similar new exemplar, compared to prior presentation of a dissimilar new exemplar. This suggests information about old exemplars other than the prototype was retrieved in categorization of the new exemplar. According to prototype theory, categorization of both prior exemplars should cause the prototype to be retrieved, and therefore equal priming should have been observed for both stimuli, which was not the case. Similarly, all of the old exemplars were not retrieved for each categorization, as if they were then there would again be no difference in priming. (Using a second

condition where participants were required to make a perceptual judgement about the new exemplar, Malt failed to obtain a priming effect. Thus a perceptual enhancement account of the priming seems unlikely.) People are sensitive to correlation information about different features within a category (Ashby & Gott, 1988; Ashby & Maddox, 1992; Medin, Altom, Edelson, & Freko, 1982) that would be lost if they retained only a prototype. Medin and Schwanenflugel (1981) demonstrated that in some cases participants can classify non-linearly separable stimuli at least as easily as linearly separable stimuli, a finding which cannot be explained by a prototype model. Together these findings all suggest that the representation of a category consists of memory for more than just the category prototype.

Exemplar Models of Categorization

Exemplar models (Ashby & Maddox, 1993; Brooks, 1978; Estes, 1994; Lamberts, 1994; Medin & Schaffer, 1978; Nosofsky, 1986) assume participants represent categories by storing every single stimulus encountered, together with its category label. A novel item is classified by calculating the similarity between the item and the stored examples. The notion of similarity is rather unconstrained (Goodman, 1972), and therefore problematic. The models of categorization described here overcome the criticism that any two items can be similar or (dissimilar) in an infinite number of ways by specifying over what features or dimensions similarity is to be considered. With perceptual stimuli, the implementation of similarity in models differs: some models use a spatial metaphor (e.g., Ashby & Perrin, 1988; Ashby & Townsend, 1986; Medin & Schaffer, 1978; Nosofsky, 1984; Nosofsky, 1986), and others use feature matching (Tversky, 1977; Tversky & Gati, 1982).

Whichever model of similarity is used, a stimulus can be thought of as

represented as a precise point in multidimensional psychological space, called an exemplar. The distance between two points in multidimensional psychological space is related to the similarity between the two exemplars they represent, i.e., the amount of generalization between the exemplars (Carroll & Wish, 1974b; Shepard, Romney, & Nerlove, 1972). Shepard (1958) showed that the idea that distance in psychological space could be related to similarity could be derived from stimulus generalization in learning theory (Hull, 1943). Nosofsky (1984) applied the proposal that distance in psychological space could be related to generalization to the problem of similarity in categorization. The structure of psychological space is often determined by multidimensional scaling (MDS), whereby pair wise similarity judgments or identification confusion data uniquely determine the relative co-ordinates of the stimuli in the space (Shepard, 1974; 1980). To classify a stimulus a participant derives its similarity to each of the stored exemplars. Often, the probability of classifying a stimulus as a member of a particular category is a function of the summed similarity to all the category's exemplars and the summed similarity to all of the possible categories' exemplars. Normally Luce's (1959) choice rule is used to map the summed similarities for each category onto the probability of responding with each category label.

Nosofsky's (1986) exemplar model is now described. Other exemplar models are described in relation to this model, as many of them are either restricted versions of this model, or are closely related to it.

The Generalized Context Model

The generalized context model (GCM, Nosofsky, 1986) is an extension of the similarity choice model for predicting identification confusion data (Luce, 1963; Shepard, 1957; Smith, 1980; Townsend & Landon, 1983).

Let $X_k=\{x^n; n=1, \ldots, N_k\}$ be the set of stored category C_k examples. x^n is a vector in multidimensional psychological space (where the superscript n denotes the n^{th} trial and is derived from a MDS procedure). The probability that the stimulus x^n is classified in category C_k (where different values of the subscript k denote different categories) is given by

$$P(C_k|\mathbf{x}^n) = \frac{\beta_k h_k(\mathbf{x}^n)}{\sum_{i=1}^{K} \beta_i h_i(\mathbf{x}^n)} \quad (1)$$

where β_k is a response bias, and $h_k(x^n)$ is the summed similarity between x^n and every stored category C_k example:

$$h_k(\mathbf{x}) = \sum_{n=1}^{N_k} \exp\left(-c.d(\mathbf{x},\mathbf{x}^n)^q\right) \quad (2)$$

where q=1 yields an exponential function, and q=2 yields a Gaussian function; $d(x,x^n)$ is a measure of the psychological distance from x to x^n. The non-negative parameter c scales the psychological space and can be interpreted as a measure of the overall stimulus discriminability, or as the amount of generalization between stimuli.

The psychological distance $d(x,x^n)$ is computed using a weighted Minkowski r-metric in a d-dimensional space:

$$d(\mathbf{x},\mathbf{x}^n) = \left[\sum_{i=1}^{d} w_i \left|x_i - x_i^n\right|^r\right]^{1/r} \quad (3)$$

where w_i is the proportion of attention allocated to dimension i. The exponent r defines the distance metric: the value r=1 produces the city block metric, r=2 produces the Euclidean metric. A Euclidean distance metric is most appropriate for stimuli with integral dimensions, and a city block metric for those with separable dimensions (Garner, 1974; Nosofsky, 1987).

The Relationships Between Exemplar Models of Classification

The Context Model

Nosofsky (1986) has demonstrated that Medin and

Schaffer's (1978) context model, hereafter CM, is a special case of the GCM where the dimensions are binary and q=r=1. Intuitively, in the context model, the similarity between two exemplars is given by s raised to the power of the number of features the two exemplars differ on, where s is the similarity between two exemplars differing on only one feature, and 0<s<1.

The Weighted Ratio Model. Lambert's (1994) weighted ratio model has a different definition of similarity. In the case where all dimension weights are the same, the summed similarity of exemplar $\underset{\sim}{x}^n$ to all the members of C_k is given by

$$h_k(\mathbf{x}) = \sum_{n=1}^{N_k} \frac{(1-t)CF(\mathbf{x},\mathbf{x}^n)}{(1-t)CF(\mathbf{x},\mathbf{x}^n) + t.DF(\mathbf{x},\mathbf{x}^n)} \tag{4}$$

where $CF(\underset{\sim}{x},\underset{\sim}{x}^n)$ is the number of dimensions or features $\underset{\sim}{x}$ and $\underset{\sim}{x}^n$ have in common (the number of common features), $DF(\underset{\sim}{x},\underset{\sim}{x}^n)$ is the number of dimensions $\underset{\sim}{x}$ and $\underset{\sim}{x}^n$ differ on (the number of different features) and t is the relative weight such that 0<t<1. This similarity function is effectively a weighted ratio of the number of common features over the number of common plus number of different features. It differs from the similarity functions for the CM and GCM in that similarity in the CM and GCM is only a function of the number of different features, rather than the number of similar and the number of different features. If the number of dimensions, r, in a given experiment does not vary then the number of different features is a function of the number of common features and same the summed similarity of $\underset{\sim}{x}$ to all the members of C_k becomes

$$h_k(\boldsymbol{x}) = \sum_{n=1}^{N_k} \frac{(1-t)\left(r - d(\boldsymbol{x},\boldsymbol{x}^n)\right)}{(1-t)\left(r - d(\boldsymbol{x},\boldsymbol{x}^n)\right) + t\left(d(\boldsymbol{x},\boldsymbol{x}^n)\right)} \tag{5}$$

where $d(\underset{\sim}{x},\underset{\sim}{x}^n)$ is distance given by the city block metric in a psychological space with a binary dimension representing the presence or absence of each feature.

The Exemplar-Similarity Model. Estes' (1994) exemplar-similarity model is

an extension of Medin and Schaffer's (1978) CM. The probability that a stimulus is classified as belonging to a particular category is a function of the stimulus's similarity to known category exemplars as in the GCM (Equation 1). The similarity function is exponential, as for the CM and the GCM with q=1 (Equation 2). As dimensions are binary, the similarity between two exemplars on a particular dimension is either unity or s, where $0>s>1$.

At the start of category learning, when there are no category exemplars in memory then the parameter s_0 is used to represent the average similarity of any stimulus to any category. Without this assumption (or with s_0=0) then there is no effect of repeating trials on categorization performance (see Appendix 3.1 of Estes, 1994). Heit (1994) has demonstrated that a model that explains the effects of prior knowledge by treating prior knowledge as initial exemplars in the new concepts provides a good fit to empirical data, lending support to this assumption.

Each presented exemplar has a probability, p, of being encoded. Each encoded exemplar has a probability, 1-α, of being forgotten on a particular trial. (Typically once α has been included in the model, including p has little effect on the goodness of fit to empirical data.) Estes included the α parameter to account for re-learning at virtually identical rates after either an early or late change in category assignments for re-presented exemplars (Estes, 1989).

The Deterministic Exemplar Model. The deterministic exemplar model (Ashby & Maddox, 1993), hereafter DEM, contains the GCM as a special case. There are two main differences between the exemplar models discussed here, and the parametric models discussed later. First, they make different assumptions about how a category is represented. Second, they make different assumptions about how this information is integrated when making a categorization decision. The DEM has the

same representational assumptions as the exemplar models, but, rather than using a probabilistic decision rule, a deterministic decision rule and a decision bound is used:

$$\text{respond A if } \log\left(h_A(\mathbf{x})\right) - \log\left(h_B(\mathbf{x})\right) < \delta + e\text{; otherwise respond B} \tag{6}$$

where the participant is biased against category B if $\delta<0$, e is the noise in the decision, and $h_A(x)$ is the summed similarity of exemplar x to each exemplar of category A (Equation 2).

Ashby and Maddox (1993) show that using this noisy deterministic decision rule is equivalent to using this probabilistic decision rule

$$P\left(C_A|\mathbf{x}^n\right) = \frac{\beta\left(h_A(\mathbf{x}^n)\right)^\gamma}{\beta\left(h_A(\mathbf{x}^n)\right)^\gamma + (1-\beta)\left(h_B(\mathbf{x}^n)\right)^\gamma} \tag{7}$$

where

$$\gamma = \frac{\pi}{\sqrt{3}\sigma} \text{ and } \beta = \frac{e^\gamma}{1+e^\gamma}$$

and the noise, e, has a logistic distribution of mean 0 and variance σ. The DEM can be thought of as the GCM with one additional parameter, γ. Depending on the value of γ responding is either more or less variable than the GCM predicts. Nosofsky (1991) also produced a deterministic exemplar model, but it does not contain the GCM as a special case – Luce's (1959) choice rule is abandoned in favor of a winner take all deterministic rule, where the category associated with the highest summed similarity is always given in response.

Parametric Models of Classification

An alternative approach to including information other than the category prototype in the representation of a category is given by parametric models of classification. Parametric approaches to categorization include general recognition

theory (Ashby & Townsend, 1986), prototype models (e.g., Homa et al., 1981; Posner & Keele, 1968; Posner & Keele, 1970; Reed, 1972; Rosch, 1973; Rosch et al., 1976), general linear classifiers (e.g., Medin & Schwanenflugel, 1981; Morrison, 1990; Nilsson, 1965; Townsend & Landon, 1983), optimal decision rules (e.g., Fukunaga, 1972; Green & Swets, 1966; Noreen, 1981; Townsend & Landon, 1983) and the category density model (Fried & Holyoak, 1984). Parametric approaches assume that a specific functional form can represent the density of category members in multidimensional psychological space. The form chosen depends on the particular model, but is often a multivariate normal distribution.

There are three main reasons why a normal distribution is often used to approximate natural categories: (a) Normally distributed categories share several features in common with natural categories. Both contain a very large number of potential exemplars (although this is also true of almost any other probability density function). The dimensions of both natural and normally distributed categories are continuous-valued. Many natural categories overlap, as normally distributed categories can. Many researchers (e.g., Ashby, 1992) have used these common properties to justify their choice of category structure. (b) Participants enter a category learning task assuming the categories are roughly uni-modal and symmetric. Flannagan, Fried and Holyoak (1986) have shown that normally distributed categories can be learned faster than multi-modal categories, and that in the early stages of learning a multi-modal category participants respond as if the category were uni-modal. Participants can be facilitated in learning a multi-modal category if the previously learned structure was not normally distributed. These findings do not, however, rule out other uni-modal representations. (c) If memory is limited then assuming categories are multivariate normal approximation is the best

solution in terms of maximum entropy if only the mean and covariance matrix are known (Myung, 1994). These three reasons certainly do not compel the use of a normal distribution as a category representation, but this choice is commonly used. Once the form used to approximate the category is chosen, building a representation of the category reduces to estimating the functional form's free parameters. The value of the category density function at a particular point in psychological space for each category is used to compute the classification response for the stimulus represented by that particular point.

Ashby's (1986) general recognition theory is now described. Other parametric models of categorization are described with reference to this model, as with the exception of Fried and Holyoak's (1984) category density model, they are all special cases of general recognition theory.

General Recognition Theory

General recognition theory (GRT, Ashby & Townsend, 1986) is a multidimensional generalization of signal detection theory (Green & Swets, 1966; Swets, Tanner, & Birdsall, 1961). Thus a stimulus, $\underset{\sim}{x}^{n}$, is represented by a vector in multidimensional psychological space, $\underset{\sim}{x}^{n}_{p}$, where the subscript p means perceived. GRT assumes there to be noise in the perceptual system, $\underset{\sim}{e}^{n}_{p}$, and therefore repeated presentation of the same stimulus, $\underset{\sim}{x}^{n}$, does not always lead to the same perceptual representation, $\underset{\sim}{x}^{n}_{p}$.

$$\mathbf{x}_p^n = \mathbf{x}^n + \mathbf{e}_p^n \tag{8}$$

Typically, the noise is assumed to be multivariate normal, with covariance matrix $\underset{\sim}{\Sigma}^{n}_{p}$, i.e., $\underset{\sim}{e}^{n}_{p}=N(\underset{\sim}{0}, \underset{\sim}{\Sigma}^{n}_{p})$. The perceptual effects of each example of a category can be represented by a multivariate normal distribution

$$p\left(\mathbf{x}^n\right) = \mathrm{N}\left(\mathbf{x}^n, \Sigma_p^n\right) \tag{9}$$

The perceptual representation of a category is a probability mixture of the individual example distributions. If $\underline{X_k}=\{\underline{\mathbf{x}}^{\underline{n}}; \underline{n}=1, \ldots, \underline{N_k}\}$ is the set of stored category $\underline{C_k}$ examples then the probability density function associated with this category is given by

$$p(\mathbf{x}|C_k) = \sum_{n=1}^{N_k} P(\mathbf{x}^n|C_k)\mathrm{N}(\mathbf{x}^n, \Sigma_p^n) \tag{10}$$

where $P(\underline{\mathbf{x}}^{\underline{n}} \mid \underline{C_k})$ is the probability that stimulus $\underline{\mathbf{x}}^{\underline{n}}$ is presented as a member of category $\underline{C_k}$.

By constraining the covariance matrix $\underline{\underline{\Sigma}}^{\underline{n}}_{\underline{p}}$ special cases of GRT can be derived. The stimulus invariant GRT assumes that $\underline{\underline{\Sigma}}^{\underline{n}}_{\underline{p}}=\underline{\underline{\Sigma}}_{\underline{p}}$ for all $\underline{n}$. The un-correlated GRT assumes $\underline{\underline{\Sigma}}_{\underline{p}}$ is diagonal and the simple GRT assumes that $\underline{\underline{\Sigma}}_{\underline{p}}=\underline{\sigma^2}_{\underline{p}}\underline{\mathbf{I}}$. (Ashby & Maddox, 1993.)

Once the percept $\underline{\underline{\mathbf{x}}}^{\underline{n}}_{\underline{p}}$ of a stimulus is formed, response selection in the GRT is a deterministic process – the category for which the probability of the data, given the category, is maximized is chosen as category label. It is assumed that a participant divides perceptual space into distinct category regions and responds according to which region the percept falls into. The border of each of these regions is called a decision bound. Stimuli that fall on the decision bound are therefore equally likely to be classified into either category. The decision bound can be computed as

$$d_{\{k,l\}}(\mathbf{x}) = P(C_k)p(\mathbf{x}|C_k) - P(C_l)p(\mathbf{x}|C_l) \tag{11}$$

where $\underline{d}_{\{\underline{k},\underline{l}\}}(\underline{\mathbf{x}})$ is the decision bound between category $\underline{C_k}$ and category $\underline{C_l}$. The value of the decision bound function determines which region of space the percept falls into:

$$\text{if } d_{\{k,l\}}(\mathbf{x}) \begin{cases} > 0 \text{ respond } C_k \\ = 0 \text{ guess} \\ < 0 \text{ respond } C_l \end{cases} \tag{12}$$

As this decision bound is the one that maximizes overall categorization accuracy this decision bound is called the GRT optimal classifier (Duda & Hart, 1973). Intuitively, this corresponds to classifying a stimulus into the category it is most likely to belong to, according to the inferred category probability density functions. The perceptual noise turns this deterministic process into a probabilistic one.

If the examples are normally distributed within the category then the category density function $P(\underset{\sim}{x} \mid C_k)$ can be rewritten as

$$p(\mathbf{x}|C_k) = \mathrm{N}(\mathbf{x};\mu_{pk},\Sigma_{pk}) \tag{13}$$

where μ_{pk} and Σ_{pk} denote the perceived category C_k mean vector and covariance matrix respectively. This model is called the normal or Gaussian GRT. Ashby (1992) makes the strong assumption that a category can be represented by a multivariate normal probability density function even when the true example distribution within the category is not normally distributed. A subject using a normal GRT classifier will always have a quadratic decision bound (unless the two category covariance matrices are equal, then the boundary will be linear).

The version of the GRT used in this thesis is a further constrained version of the normal GRT. The further constrain is that the variance-covariance matrix Σ_{pk} is constrained to be diagonal and to have equal variance for each dimension.

$$\Sigma_{pk} = \sigma_{pk}^2 \mathbf{I} \tag{14}$$

Thus, only two free parameters must be estimated for each category, the mean and the variance.

Normal GRT is very similar to decision bound theory. Normal GRT and decision bound theory predict the same categorization decision and accuracy for every stimulus. However, participants using GRT are assumed to estimate

parameters for the probability density function for each category, which uniquely determine the decision bound's parameters. In decision bound theory participants are assumed to directly estimate the free parameters of the decision bound.

Relationships Between Parametric Models of Categorization

Prototype Models. Prototype models (e.g., Homa et al., 1981; Posner & Keele, 1968; Posner & Keele, 1970; Reed, 1972; Rosch, 1973; Rosch et al., 1976) also use a multidimensional psychological space. The prototype is the central tendency of the category, and corresponds to the category mean in GRT. A stimulus is categorized as belonging the category with the most similar prototype, where similarity corresponds to the distance between the stimulus and the prototype in multidimensional psychological space.

In a binary classification prototype models classify as linear decision bound models, where the bound's slope and intercept are completely determined by the two category means. Any point on the bound is equidistant from both category means. Prototype models are a special case of the Gaussian GRT where the co-variance matrix for the each category is constrained to have zero co-variances and equal variance for each dimension and each category:

$$\Sigma_{pk} = \sigma_p^2 \mathbf{I} \tag{15}$$

The General Linear Classifier. The general linear classifier (GLC) (e.g., Medin & Schwanenflugel, 1981; Morrison, 1990; Nilsson, 1965; Townsend & Landon, 1983), also uses a linear decision bound. Unlike the prototype models, the slope and intercept of the decision bound are not constrained, and are varied to provide optimal categorization performance. In other words, participants are assumed to estimate the slope and intercept of the decision bound directly, rather than deriving them from category probability density functions. However, the

optimal bound is that predicted by GRT.

The General Quadratic Classifier. The general quadratic classifier (GQC) (e.g., Ashby, 1992; Ashby & Maddox, 1992) uses a quadratic decision bound. When two categories are normally distributed, but have non-equal covariance matrices the GQC is the optimal decision bound. The GQC makes identical categorization predictions to the Gaussian GRT, and contains the GLC as a special case. As for the general linear classifier, participants are assumed to estimate the free parameters of the decision bound directly, but the optimal values are those predicted by GRT.

Optimal Decision Rules. The optimal decision rule or bound (e.g., Fukunaga, 1972; Green & Swets, 1966; Noreen, 1981; Townsend & Landon, 1983) is the one that maximizes accuracy of categorization, and is therefore not necessarily linear or quadratic. In the two category case every stimulus represented by a point on the decision bound is equally likely to belong to either category. Unless assumptions are made about the density function for each category then it is hard to say anything about the shape of the bound. If normality is assumed then categorization accuracy is as the Gaussian GRT or the GQC.

The Category Density Model. Fried and Holyoak's (1984) category density model uses the same representational assumptions as Gaussian GRT, with the caveat that category covariance matrices are constrained to be diagonal. (A surface of equi-probability for a category is therefore constrained to be a sphere.) In addition, the model uses a tally of the frequency of occurrence of each category. Luce's (1959) choice rule is used to calculate the probability that exemplar $\underset{\sim}{x}^n$ is considered to be a member of category C_k

$$P\left(C_k \middle| x^n\right) = \frac{\beta_k \psi_k\left(x^n\right)}{\sum_{i=1}^{K} \beta_i \psi_i\left(x^n\right)} \tag{16}$$

where β_i is a decision bias for each category, K is the number of alternative categories and Bayes' theorem is used to give the subjective probability that the decision maker considers item x^n to be a member of category C_k

$$\psi_k(x^n) = \frac{P(x^n|C_k)P(C_k)}{\sum_{i=1}^{K} P(x^n|C_i)P(C_i)} \tag{17}$$

where $P(x^n \mid C_k)$ is the subjective conditional probability of item x_n occurring given category C_k, and $P(C_k)$ is the subjective prior probability of C_k. Fried and Holyoak refer to Equations 16 and 17 together as the relative likelihood decision rule. (Note how similar this decision rule is to that of the GCM: the similarity to a category function has been replaced by a probability of belonging to a category function.)

The category density model was proposed as a model of category learning. An algorithm is provided whereby learning may take place in the absence of feedback, and even in the absence of knowledge of the number of categories to be learnt. With feedback, the first exemplar of a category provides the category mean. The first and second exemplars together with initial parameters representing the learner's prior expectations of the variance of the category along each dimension are used to estimate the diagonal elements of the covariance matrix. In the absence of feedback category, means and variances are estimated after the first s exemplars, where $s>K$. (s represents the size of the short term memory buffer.) Initially s groups are defined, each containing one of the s exemplars so far. A clustering algorithm is used to divide the s exemplars into K groups, by repeatedly grouping together the groups with the closest centroids. The mean of each resulting group then becomes the mean of each category, and the initial variances for each dimension for each category are constructed by pooling an arbitrarily large value with the variance of each group (as in the feedback condition).

All that remains now to complete the model is to give an account of how the means, variances, and frequencies for each category are updated with each new exemplar. In the feedback condition, the value of $\underline{\Psi_i}$ at time $\underline{t}$,

$$\psi_{i,t} = \begin{cases} 1, \text{ if } \mathbf{x}^n \text{ is labelled as a member of } C_k \\ 0, \text{ otherwise} \end{cases} \quad (18)$$

In the no feedback condition Equation 17 for time $\underline{t}$-1 is used to calculate $\underline{\Psi_{i,t}}$. Now depending on the magnitude of $\underline{\Psi_{i,t}}$ the mean and covariance vectors for each category are updated using standard procedures for revising running means, frequencies and variances (Raiffa & Schlaifer, 1961). Two additional bias parameters are included in the model. One reflects the degree to which participants believe each category is equally frequent, and the other reflects the degree to which participants believe the variances on each dimension are equal across categories.

Empirical Evidence for Exemplar and Parametric Models of Categorization

In the introduction to this chapter, evidence was presented that could not be explained by a simple prototype theory. The discussion here begins with how, in principle, exemplar and parametric accounts of categorization are able to account for these data. Work is then reviewed which contrast the fits of exemplar and parametric models to human categorization performance.

Three main empirical results were mentioned in the introduction. First mentioned was the typicality effect, where some members of a category are rated as better or worse examples of the category than others. In applying an exemplar model it is assumed typicality judgments and associated measures such as verification time are based on the summed similarity of the probe exemplar with all members of the category. Exemplar accounts are able to accommodate the typicality effect result because of the way MDS is used to derive the stimulus space. Items that are

particularly distinctive and not confusable are placed in the extreme regions of multidimensional psychological space. Thus atypical exemplars are of minimal summed similarity to all the other category members because they are distant from the other category members. Therefore will be rated as less typical, than typical exemplars which will be in the central regions of each category. For parametric models the probability of category membership is derived from the category density function. For reaction time measures, the distance of the exemplar from category boundaries is assumed to determine categorization latency (Ashby, Boynton, & Lee, 1994). Because distinctive atypical exemplars are less close together in physical space, the category density associated with that region of space will be low, and therefore parametric models predict that these items will be rated as less typical than items from areas of space associated with a higher category density.

The second type of finding described was cases where there is evidence that information from a specific, old exemplar influences performance. Exemplar models can, not surprisingly, accommodate all of these cases. Parametric models struggle to account for many of these specific exemplar effects. Two problematic experiments are discussed here. Homa, Stirling and Trepel (1981) showed that old exemplars can be classified more accurately than new exemplars equidistant from the prototype. If the category probability density function used is, for example, normal, then the probability of category membership is equal for the old and new exemplars for Homa, Stirling and Trepel's (1981) stimulus structure. Thus without resorting to a complicated probability density function a parametric account is unable to account for these results. Perhaps less challenging for the parametric models is Malt's (1989) priming experiment which provided evidence for the retrieval of individual exemplars. The facilitation of categorization of an old exemplar after categorization

of a new exemplar could possibly be explained by claiming that looking up or calculating the value of the PDF for one exemplar will facilitate the process for a similar exemplar more than a dissimilar exemplar. However this explanation is not very satisfactory. (It is interesting to note that both these problematic experiments involved repeated presentation of a small set of exemplars. This point is taken up below.)

The final finding discussed was that people are sensitive to the correlation between features or dimensions of stimuli. An exemplar model is able to predict this because memory for each individual exemplar maintains the correlation information in the representation. Many parametric models are also able to account for the results. Those models, where the covariance matrix used to represent each category (e.g., normal GRT) is not constrained to have equal elements on the diagonal, and non-zero covariances, maintain the correlation information in the covariance matrix for each category. Thus surfaces of equi-probability will be ellipsoids, with the correlation indicated by the orientation of the major axis.

To summarize so far, exemplar and parametric models are able to provide accounts of the main findings problematic for classical theory and prototype models. Parametric models struggle to account for data supporting episodic retrieval of specific exemplars. Examination of one popular category structure reveals that exemplar models only outperform prototype models because they reproduce accurate performance on old training items better than parametric models (Smith & Minda, 2000). In other words, the assumption that participants access memories of each training example is not necessary to explain their generalization to novel test items. When prototype models of categorization are granted with the ability to predict performance on training items, there is no difference between the fits of the two

models. In other words, ignoring performance on training items, both exemplar and distributional models provide equally good accounts of the data. Comparison of parametric and exemplar models fits to other empirical data will now be considered.

Maddox and Ashby (1993) compared the performance of the DEM, the GCM and decision bound models on a variety of data sets. The first five data sets are taken from Ashby and Maddox (1992). These data sets all involve normally distributed categories and are illustrated in Figure 1 together with examples of the stimuli used in the experiments. (Because of the very large number of exemplars participants were exposed to, MDS solutions could not be derived for the stimuli, as is normal in the fitting of the GCM.) Participants were asked to classify rectangles that could vary in height and width. These dimensions have been shown to be integral (e.g., Garner, 1974; Wiener-Erlich, 1978). Different participants classified circles that could vary in size, with diameters that could vary in orientation. These dimensions have been found to be separable (e.g., Garner & Felfoldy, 1970; Shepard, 1964).

The category structures for the first two data sets involved two categories with identical covariance matrices. The variance elements were all identical and the co-variance elements were all zero, thus the optimal decision bound is linear. Participants were given a large number of trials. The models were fit to single participant data. The DEM provided the best fit for 3 participants. The GCM provided the best fit for 1 participant. The remaining 8 participants' data were best fit by the decision bound model (the general linear classifier), although only slightly better than the DEM.

The next three data sets also involved two normally distributed categories, but with unequal co-variance matrices. The optimal decision bounds will therefore be non-linear and are in fact quadratic (Ashby & Gott, 1988; Morrison. 1990). 23 of

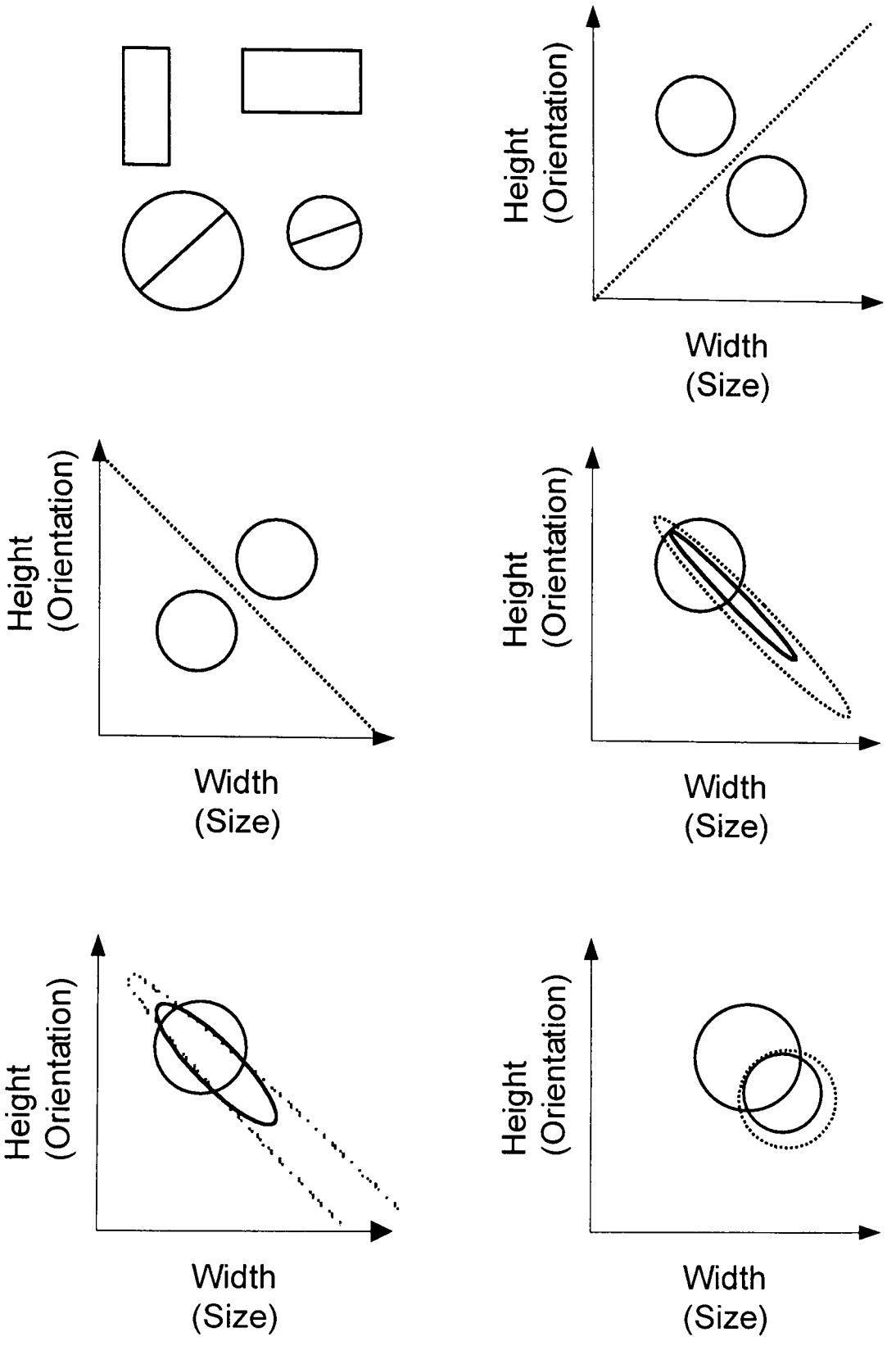


Figure 1. The stimuli and the five category structures used by Ashby & Maddox (1992). The top left panel shows two examples of rectangle stimuli and two examples of the circular stimuli. The remaining panels show the contours of equal likelihood (solid lines) and optimal decision bounds (dashed lines) for five category structures used. For the circular stimuli the labels height and width should be replaced with orientation and size respectively.

the 24 participants tested exceeded the maximum accuracy possible predicted from a linear decision bound. Most participants failed to reach the level of performance that would be predicted if they were perfectly using the optimal quadratic decision bound. The decision bound model, the DEM and the GCM were applied to data from the first and last training sessions. For the first training session the decision bound model (the general quadratic classifier) best fit data from 16 participants, the DEM best fit in 4 participants, and the GCM best fit in the remaining 4 participants. For data from the last training session the decision bound model best fit the data from 16 participants, the DEM from 5 participants and the GCM from the remaining 3 participants.

So, modeling individual participant data from participants tested with normally distributed categories, the GCM was outperformed by the DEM, which in turn was out performed by the relevant decision bound model. Maddox and Ashby (1993) also compared the GCM, the DEM and the decision bound model's fit of data when participants categorized examples from non-normally distributed categories using Nosofsky's (1986; 1989) data. The stimuli are the same as the circles with diameters as described before, except the bottom half of the stimulus is missing. The category structure used for these data sets is shown in Figures 2 and 3.

The first data set with non-normally distributed categories (Nosofsky, 1986) consists of data from two participants who participated in a large number of identification trials for the 16 possible exemplars, so that a MDS solution of the psychological space representation could be calculated for each participant for the data set. Each participant then took part in four categorization sessions. Each session began with a training phase, where participants were presented with repeated examples of half of the possible exemplars (half of these assigned to each category).

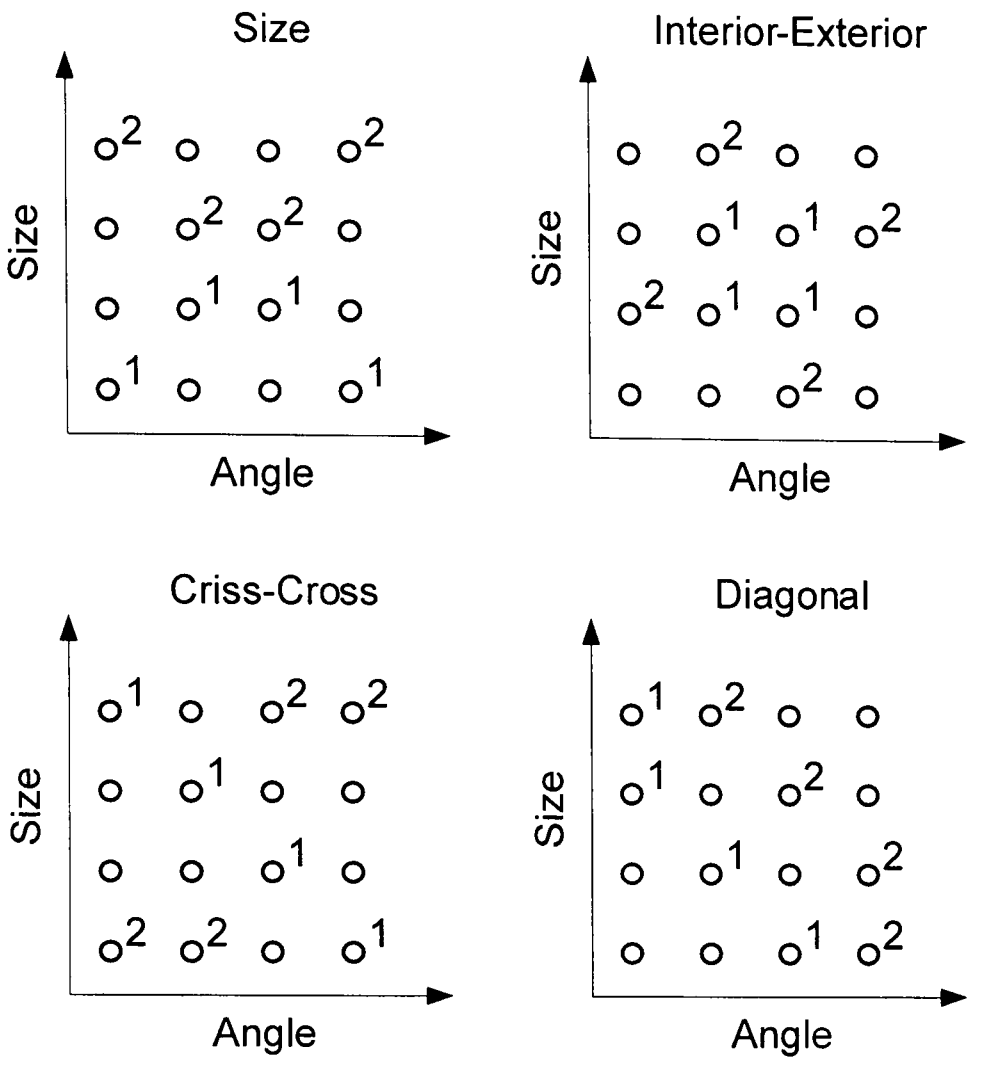


Figure 2. The four category structures used by Nosofsky (1986). The circles correspond to exemplars. Numbered circles are training exemplars, and the numbers correspond to the category assignment. The remaining exemplars were seen only during the transfer phase.

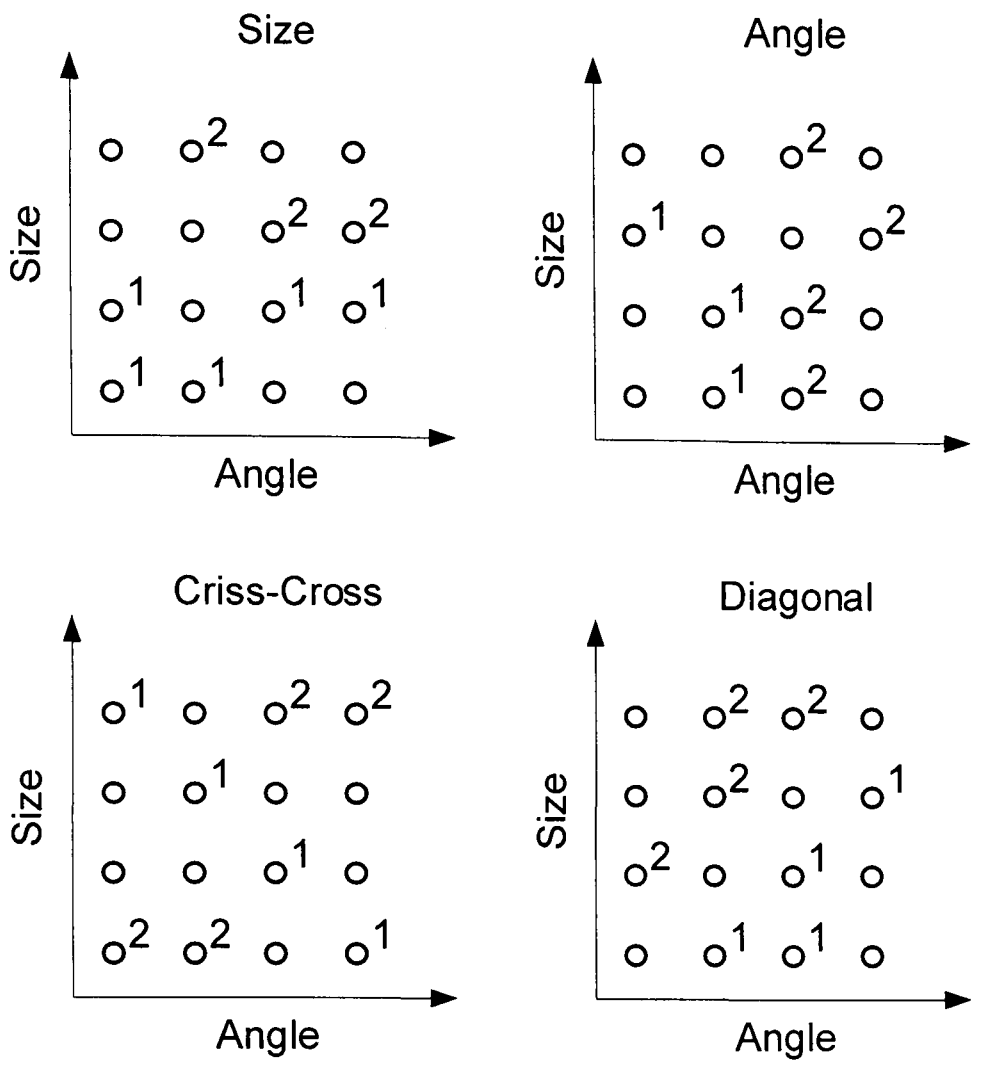


Figure 3. The four category structures used by Nosofsky (1989). The circles correspond to exemplars. Numbered circles are training exemplars, and the numbers correspond to the category assignment. The remaining exemplars were seen only during the transfer phase.

In the subsequent transfer task all exemplars were repeatedly, sequentially presented for categorization with feedback given only when a training exemplar was presented. There were four conditions: two conditions (size and diagonal) where the optimal decision bound is approximately linear and two conditions the optimal decision bound is non-linear. Maddox and Ashby (1993) fitted the DEM, the GCM and two decision bound models (the general linear classifier and the general quadratic classifier). In the two conditions when the best fitting decision bound was approximately linear there was little difference between the fits of the two exemplar and the two decision bound models. For the conditions where the best fitting decision bound is non-linear the general quadratic classifier fitted the data substantially better than the two exemplar models. Ashby and Lee (1992) fitted the optimal decision bound model, and although it did perform better than the GCM it performed more poorly than the general quadratic classifier, suggesting participants used a non-optimal classification strategy.

The second data set from non-normally distributed categories (Nosofsky, 1989) differs from the above set in four ways. First, a much larger number of participants was run. Separate participants were used to derive the MDS solution. The two dimensional MDS solutions looked much like the arrangement of the stimuli in physical space shown in (Figure 3). Four new sets of participants took part in one condition of the categorization stage of the experiment, each set in a different condition. Second, data was averaged across participants, a procedure which disadvantages the decision bound models (Maddox, 1999). Third, participants were given a smaller amount of training before the transfer phase. Finally, one of the conditions where the best fitting decision bound was non-linear was replaced with a new condition where the best fitting decision bound was linear.

As a test of the attention-optimization hypothesis (Nosofsky, 1986) three versions of the GCM were fitted to this data set (see also Nosofsky, 1987; Nosofsky, 1989; Nosofsky, 1991). The idea is that selective attention on dimensions in the psychological space will operate to optimize categorization performance. One of the versions of the GCM fitted is unconstrained, as described above. The other two versions are constrained in some way. The equal attention GCM assumes that the values of w_i in Equation 3 are equal for all d dimensions. The equal bias GCM assumes that the values of β_k are equal for all k categories. The unconstrained GCM could not be rejected for any of the categorization conditions. The equal attention GCM fitted the data significantly worse in all four conditions. Some critical transfer stimuli were designed so changes in the relative weighting of the dimensions would change their probability of assignment to a given category. Observation showed that the equal attention GCM failed to account for the probability, but the unconstrained GCM did. These results support the idea that the attention weightings for the dimensions was different in the identification and categorization tasks. The equal bias GCM fitted the data almost as well as the unconstrained GCM, suggesting that these results could not be explained by response bias. Lamberts and Chong (1999) provide further evidence that this phenomena is indeed an attention shift, by showing that categorization performance changes as a result of verbal instructions directing attention to particular stimulus dimensions.

The importance of attention shifts may seem problematic for parametric models, given that they do not incorporate attentional parameters to modify the perceptual space. However, even without assuming that the perceptual space is modified across the four conditions, decision bound models can account for the results. Maddox and Ashby (1993) also modeled the data, fitting the DEM, the GCM

and two decision bound models (the general linear classifier and the general quadratic classifier). The GCM provided the best fit in two conditions (criss-cross and diagonal), and the general linear classifier provided the best fit in the other two conditions (size and angle). It should be noted that the difference between the goodness of fits is very small. Further, the GCM fits the averaged participant data better than the single participant data. That the general linear classifier better fits the two conditions where only one dimension is needed to make the categorization than the GCM suggests that shifts in decision bounds may be more important than shifts in selective attention. Certainly though, the parametric model is able to account for the "attention shift" phenomena better than the exemplar model.

To summarize the results from data sets from non-normally distributed categories, when modeling individual participants' data after extensive training the decision bound models provided excellent accounts of the data. However when modeling data collapsed across participants after a shorter period of training the difference between the fits of exemplar and decision bound models was smaller. Overall we have seen that exemplar models like the GCM and parametric decision bound models are able to provide a very good account of empirical data. Decision bound models outperformed the GCM when category structures were normally distributed. However with non-normally distributed categories where a smaller number of exemplars were used, the both models performed about equally.

The Relationship between Exemplar and Parametric Models of Classification

Exemplar and parametric or distributional models can be though of as lying at opposite ends of a continuum of finite mixture models, where the number of distributions used to represent a category varies from one, as in GRT, to the number of examples of that category, as in the GCM (Ashby & Alfonso-Reese, 1995;

Rosseel, 1996). Also contained in this continuum are back propagation networks with sigmoidal activation functions (Rumelhart, Hinton, & Williams, 1986) and radial basis functions (Moody & Darken, 1989). With small numbers of hidden units (and hence small numbers of free parameters in relation to the size of the data to be modeled), neural networks are analogous to distributional models, because they can only learn data with a particular distributional structure. But if the number of hidden units is large in relation to the amount of data to be learned, then the neural network becomes analogous to an exemplar model, in that any data set can be modeled, whatever its structure, simply learning each piece of data (each example) by rote.

The relationship between exemplar and parametric models can be described formally using Rosseel's (1998) mixture model of category representation. The model makes three assumptions. First, following GRT, it is assumed that the presentation of the same stimulus does not always lead to the same perceptual effect. The perception of a stimulus is represented as a random vector in multidimensional psychological space

$$\mathbf{x}_p^n = \mathbf{x}^n + \mathbf{e}_p^n \tag{19}$$

where $\underset{\sim}{e}^n_p$ is a random vector with zero mean representing perceptual noise. $\underset{\sim}{e}^n_p$ is multivariate normal with zero mean and co-variance matrix $\underset{\approx}{\Sigma}^n_p$. Normally the noise is assumed to be stimulus invariant and is adequately described by $\underset{\approx}{\Sigma}_p$.

Second, the probability density function for the whole set of exemplars for all $\underline{K}$ categories is modeled as a finite mixture distribution (McLachlan & Basford, 1988; Titterington, 1984):

$$p(\mathbf{x}) = \sum_{j=1}^{J} P(j)p(\mathbf{x}|j) \tag{20}$$

where $\underline{J}$ is the number of mixture components used in the mixture model, $\underline{P}(j)$

denotes the mixture proportions and satisfy the constraint

$$\sum_{j=1}^{J} P(j) = 1 \text{ and } 0 \le P(j) \le 1, \tag{21}$$

and $p(\underline{x} \mid j)$ is a multivariate normal distribution with mean $\underline{\mu}_j$ and variance $\underline{\Sigma}_j$.

Third, this probability density function is shared by each of the K categories. Category C_k is modeled by the same set of mixture components $p(\underline{x} \mid j)$ as the unconditional mixture distribution $p(\underline{x})$.

$$p(\boldsymbol{x}|C_k) = \sum_{j=1}^{J} P(j|k)p(\boldsymbol{x}|j) \tag{22}$$

where $P(j \mid k)$ denotes the class-conditional mixture proportions.

The category representation of the finite mixture model contains the category representations of exemplar and decision bound models (Ashby & Alfonso-Reese, 1995; Rosseel, 1996). The category representation of the Gaussian GRT is equivalent to a finite mixture model with one multivariate normal mixture component. Rosseel (1998) has shown that the category representation of the GCM using a Gaussian similarity function and a Euclidean distance metric is equivalent to a finite mixture model with multivariate normal mixture components for each category exemplar. Further, Rosseel showed that the category representation of the GCM using an exponential similarity function and a city block distance metric is equivalent to a finite mixture model with multivariate Laplacian mixture components for each category exemplar. Thus by altering the number of mixture components, J, the mixture model can take on the representation assumptions of the GCM (when J equals the total number of exemplars across all categories), or Gaussian GRT (when J equals the number of categories).

Conclusions

Illustrative evidence that prompted the replacement of the classical and

prototype theories of categorization has been presented. The two current accounts of categorization – exemplar theories and parametric or distributional theories – have been described in detail. The relationship between various instantiations of each theory have been described. Empirical evidence collected with the intent of discriminating between these two accounts has been presented. However, it was concluded that exemplar and distributional approaches both provided good accounts of the data. Finally the relationship between exemplar and distributional views was formalized using a finite mixture model framework.

Summary of Remaining Chapters

The experiments presented in Chapter 2 of this thesis investigate generalization to novel items between two categories that differ in variability. It is shown that exemplar and parametric accounts make qualitatively different predictions for the pattern of generalization. Thus the experiments are designed to discriminate between the use of exemplar based representations and distributional representations.

The experiments in Chapter 3 of this thesis challenge the assumption that participants have access to the absolute location of stimuli in physical space, an assumption common to both exemplar and parametric models of categorization. The experiments were motivated by evidence from absolute identification and magnitude estimation paradigms, which demonstrates that participants typically have poor access to absolute magnitude information. Instead participants rely upon comparisons with recent stimuli, as evident from the strong effect of preceding material demonstrated in these paradigms. Here specific sequence effects are examined in categorization – effects that are not predicted by the use of absolute magnitude information alone. A new model of categorization is developed, where

classification is based on the relative magnitude information from comparisons with immediately preceding stimuli.

In Chapter 4 the effects of categorization experience on perception are examined. Evidence of perceptual learning during exposure or categorization is reviewed. The experiments are concerned with the specific claim that new representational features may be created to facilitate novel categorizations, and that these features qualitatively alter the perception of stimuli. Such a claim is counter to the assumption implicit in existing models of categorization that the representation of stimuli is fixed.

Finally in Chapter 5 an integrated account of the experiments is presented, together with extensive suggestions for future work to investigate the claims of the integrated account.

Chapter 2

The Effect of Category Variability in Perceptual Categorization

Abstract

Two very different views have been advanced concerning how people learn categories from labeled examples. The exemplar view suggests that people store some or all examples, and categorize new items by their similarity to these stored items. What we call the distributional view suggests, instead, that people fit probability distributions using the examples from each category, and classify new items by reference to these probability distributions. A key differential prediction between these viewpoints concerns the classification of new examples precisely intermediate between the nearest examples from two categories that differ in variability. The exemplar approach, illustrated using the generalized context model (Nosofsky, 1986), predicts that the intermediate item should typically be classified in the lower variability category. By contrast the distributional approach, illustrated using the general recognition theory (Ashby & Townsend, 1986), predicts classification into the higher variability category. Neither prediction is confirmed experimentally – instead a highly variable pattern of results is found. Experiments 1 and 2 show that classifications behavior can be strongly influenced by the salience of the difference in variability between the categories. Experiments 3 and 4 show great variation between participants on the effect of increasing the difference in variability between the two categories. Neither the exemplar nor distributional viewpoints can predict the behavior of the majority of participants.

The Effect of Category Variability in Perceptual Categorization

This paper considers the accounts of classification given by two successful models of categorization. Exemplar models (e.g., Medin & Schaffer, 1978; Nosofsky, 1986) assume the categorization of a new item is based on the similarity of the new item to examples of previously encountered items stored in memory. An alternative is that probability distributions used to represent categories, and that these distributions are fitted using the encountered examples. Classification of a new item is based on the relative likelihood of belonging to each distribution. This alternative will be called the distributional approach (e.g., Ashby & Townsend, 1986). The difference between these two accounts may be illustrated with a simple example. Consider two categories (Figure 4). The examples of one category may be more variable than the examples of the other category. If an example intermediate between nearest examples of the two categories is presented it may be classified into either category. Such an example is more similar to the examples of the low variability category, as these examples will be nearer to the intermediate item in perceptual space. Exemplar accounts must therefore predict that this intermediate example should on average be classified into the lower variability category more often. However, often it is rational to classify this example as a member of the category with the larger variability, as other things being equal, it is more likely to belong to this category. The distributional account may therefore make a different prediction for the classification of the intermediate example. It is not always the case that distributional and exemplar accounts must lead to opposite categorizations in this situation, but it is certainly possible to arrange the relative variability of the pair of categories so that the models do. The experiments described in this paper are designed to investigate the basis for generalization in categorization by manipulating

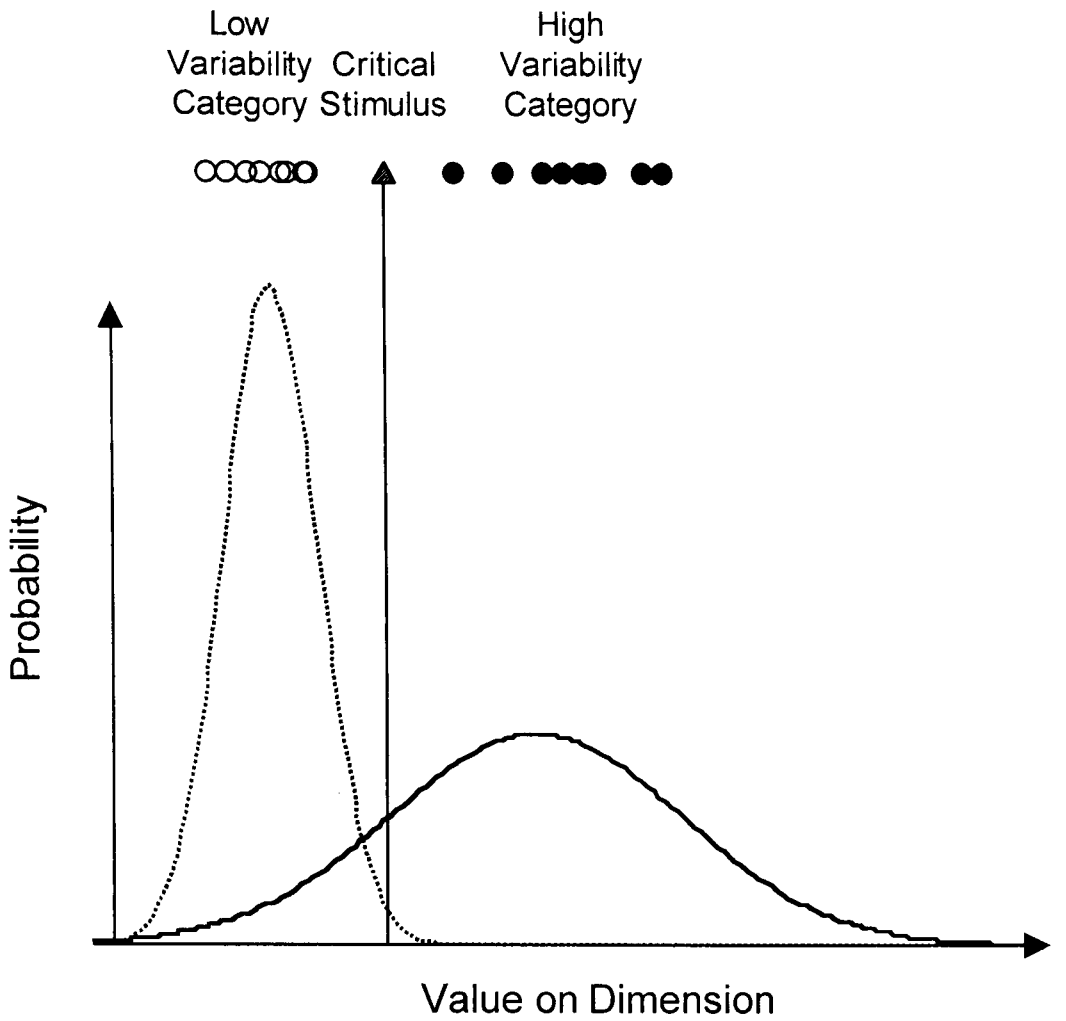


Figure 4. A one dimensional example of two categories differing in variability. The examples of the low variability category happen to take low values on the dimension (outline circles). The probability density function from which they were generated is represented by the dashed line. The examples of the high variability take high values of the dimension (filled circles). The probability density function from which they were generated is represented by the solid line. A critical example midway between the nearest examples of the two categories (triangle) is more likely to belong to the high variability category, but more is more similar to examples of the low variability category.

the relative variability of two categories in a binary categorization.

The difference between exemplar and distributional models can be viewed as an aspect of a much broader theoretical issue in cognitive science: to what extent is human learning 'lazy' or 'eager'? (Aha, 1998; Hahn & Chater, 1997; Hahn & Chater, 1998). Lazy learning involves storing input material in a relatively unprocessed form; the cognitive work required to transfer this knowledge to some new context (e.g., generalizing past experiences to a new situation) is applied only when this work needs to be done. This style of learning is 'lazy' because cognitive work is only done when strictly necessary – otherwise the learning items are simply stored. In cognitive science, lazy learning is embodied not only exemplar models of categorization (e.g., Medin & Schaffer, 1978; Nosofsky, 1986), but also in exemplar-based accounts of memory (e.g., Hintzman, 1986), case-based reasoning (e.g., Kolodner, 1993), and analogy-based models of reading and morphological processing (e.g., Glushko, 1979; Nakisa & Hahn, 1996). By contrast, eager learning involves actively attempting to extract regularities from new items, as they are encountered. The model of the regularities that has been extracted can then straightforwardly be applied to new items, as they are encountered. Eager learning methods vary between methods that involve the attempt to seek symbolic rules with which to model the incoming data (e.g., Lavrac & Dzeroski, 1993; Thagard, 1988), and those that attempt to fit incoming data to some kind of probabilistic model (e.g., Ashby & Townsend, 1986).

The effects of category variability and generalization have been addressed in two important studies, by Rips (1989) and Fried and Holyoak (1984). Rips (1989) used a binary categorization with categories of differing variability to dissociate similarity and categorization judgments. Participants were presented with sentences

giving information about an object's value on a single dimension. In one condition participants had to classify the object as a member of one of two available categories on the basis of this information alone. For example, "a circular object with a 3-inch diameter" could be classified as a pizza or a quarter. In another condition, participants were asked to choose the category the object was more similar to. Note that participants were not asked for similarity ratings between two objects as is typical in predicting classification from similarity or identification (e.g., Nosofsky, 1986), but rather gave ratings of the similarity between an object and a category. The value of the object on the selected dimension was chosen to be half way between each participant's estimate of the lowest value of the high value category, and of the highest value of the low value category. Participants were told this is how the test value they were given was derived. Rips found that similarity decisions favored the low variability category, but that categorization decisions favored the high variability category. Continuing the above example, the 3-inch diameter circular object was rated as similar to quarters, but was categorized as a pizza. The object could not be a quarter, as it was too big. Quarters have a fixed size, so even though the object may be more similar to the size of quarters it cannot be a quarter. Rips took the dissociation between similarity and categorization as evidence that categorization decisions were not based on similarity decisions.

Empirical evidence from Smith and Sloman (1994) provides a pertinent boundary condition on this dissociation. Using Rips' stimuli they were unable to replicate the dissociation. This failure to replicate Rips' dissociation is probably due to a procedural difference. In Rips' study participants were asked to talk aloud about the decisions they were making, but in Smith and Sloman's study participants simply pressed one of two keys. In a second experiment Smith and Sloman replicated Rips'

dissociation when they asked their participants to talk aloud. Analysis of the participants' protocols on each trial showed that participants were more likely to mention the feature plus corresponding reason (e.g., "quarters cannot be that big") in the categorization condition than in the similarity condition. Further, when participants mentioned the necessary feature and rule, they were almost always chose the more variable category. It seems then that Rips' dissociation of categorization and similarity is only obtained under conditions that require verbal rationalization of the categorization decision.

Rips' study leaves open the question of the effect of relative category variability in perceptual categorization, the topic of the present paper, for two reasons. First, Rips used familiar semantic categories, to encourage participants to use prior knowledge from outside the experimental context. Such knowledge is not available for the kinds of abstract perceptual category traditionally used in perceptual categorization experiments (although it may well be available for natural perceptual categories). Second, the effect that Rips describes does not seem to be robust in conditions most analogous to typical perceptual categorization task (where participants do not produce verbal protocols).

Fried and Holyoak (1984) have, though, shown that participants are sensitive to the relative variability of perceptual categories. For example, in Fried and Holyoak's (1984) Experiment 2 participants categorized black and white checkerboards. Two prototype checkerboards were generated, one prototype for each of the two categories the participants had to learn. Training examples for one category (low variability) were created by allowing each square of the prototype a small probability of changing color. For the other category (high variability) the probability was about double. As low variability categories are easier to learn a third

category (other) was introduced, to prevent subjects learning only the low variability category, and classifying all non-members as members of the high variability category. Participants classified training examples until they reached criteria. In transfer, participants categorized distortions the two standards. Significantly more transfer items were classified as belonging to the higher variability category than would be expected by chance. Further, there was a tendency to classify items into the high variability category even for patterns physically closer to the low variability category. Fried and Holyoak had predicted these findings with their category density model, and interpret these findings as support for the idea that participants use examples to induce distribution functions, and then classify according to a relative likelihood rule.

This conclusion is certainly consistent with a distributional approach, where a category probability density function is estimated for each category, based on the category's mean and variance in some multidimensional psychological space. However, Fried and Holyoak's findings are also consistent with exemplar-based categorization. Consider a transfer checkerboard with an equal number of squares in common with each of the two base prototypes. Using Fried and Holyoak's "no of squares in common" similarity estimate, such a checkerboard would be equally similar to each category prototype. Now consider though the distribution of exemplars of each category. It is very much more likely that there will be more exemplars from the high variability category near the transfer checkerboard than exemplars from the low variability category, simply because the checkerboards from the high variability category are more scattered in the space. Thus the target checkerboard will be a better example of the high variability category than the low variability category (this intuitive argument is confirmed by the simulations reported

below). Thus an exemplar-based model of categorization that assumes memory for (at least some) individual exemplars can account for the results Fried and Holyoak cite as evidence for the distributional approach[1].

A second issue regarding Fried and Holyoak's interpretation of this study is that their similarity estimate may lead to an incorrect assumption about the representation of these checkerboard stimuli. There is mounting evidence that new functional features are learned during the categorization of checkerboards (see Chapter 4, and also McLaren, 1997; Palmeri & Nosofsky, in press; Wills & McLaren, 1998). To a first approximation it seems that the largest invariant chunk of a stimulus is learned as a feature. If this is the case, then one might expect qualitatively different features to be learned for Fried and Holyoak's low and high variability features because the low variability checkerboards are more likely to contain larger invariant features than the high variability checkerboards. That is, the probability of a sizeable chunk of the low variability board remaining constant is much greater than for a chunk of equivalent size in a high variability board. This would lead to the creation of fewer and larger functional features for the low variability category than the high variability category. If this were the case, then a stimulus equally distant between the two categories may indeed be more similar to the high variability category. This is because the probability of the presence of larger chunks used to represent the low variability category is much lower than the probability of the presence of the smaller high variability chunks. In summary, categorization of a stimulus with an equal number of squares in common with the two category standards as a member of the high variability category is not

---

[1] There is evidence that people do have memory for (at least fragments of) individual examples with similar stimuli (Homa et al., 1981).

necessarily inconsistent with similarity based categorization. The closer proximity in psychological space of examples of the high variability category to the equidistant checkerboard, and / or the more likely presence of high variability category features in the equidistant checkerboard allows similarity based categorization to account for these results. In any case, without deriving a multidimensional scaling solution from pair wise similarity judgments or identification confusions, one cannot be sure of the structure of the psychological space, and therefore that these results cannot be accounted for by similarity based categorization.

What is needed a category structure that allows similarity and likelihood based categorization to be distinguished, even when memory for individual exemplars is allowed (as it is in the hugely successful exemplar models of categorization). A simple one-dimensional case of such a structure is illustrated in Figure 4. The exemplars from the high variability category are more spread out than the exemplars from the low variability category. Consider a critical item exactly half way between the nearest exemplars of each category. Formal modeling of the classification of this item is given in the next section, but for now simple argument will suffice.

If, as in exemplar models, (a) categorization is based on the comparison of the summed similarity between the critical stimulus and each stored stimulus for each category, and (b) similarity is some monotonically decreasing function of the distance between a pair of exemplars on the dimension, then the critical stimulus should be categorized as a member of the low variability category more often than the high variability category. This is because the summed distance between the critical stimulus and each example of the low variability category is smaller than the sum for the high variability category, which means the summed similarity must be

greater for the low variability category. Because the nearest examples of each category are equally distant from the critical stimulus categorization will in fact be based on the next nearest examples – which are more likely to come from the low variance category (see Figure 4), because low variance stimuli are more closely bunched near the critical stimulus.

In a distributional model, categorization will be based on the likelihood of the data, given the induced probability distribution for each category. If the presumed distribution is Gaussian (see Figure 4), then the intermediate example will typically, though not definitely, be classified as a member of the high variance category, because the tight bunching of the low variance items means that the intermediate test item would be more standard deviations from the mean of the low variance category (we assume here that the frequencies of each category are equal, or appropriately equal – in the experiments below, there is indeed no bias in favor of one category or the other). The reason why it is not certain that the critical stimulus should be categorized as a member of the high variability category more often than as a member of the low variability category is because the critical stimulus is not equidistant between the means of the two categories, (when this would always be the case). (It is worth pointing out here that if this were the case then an exemplar model would be able to predict classification of the critical example into the high variability category as this category is more likely to have the nearest neighbor, as in Fried and Holyoak's design.) Rather the critical stimulus is equidistant between the nearest neighbors of the two categories, and is therefore nearer the mean of the lower variability category. Thus the difference in variability between the two categories need be sufficiently large to counter the fact that the low variability category has the nearer mean.

This argument then predicts that, provided two categories differ sufficiently in variability, the exemplar and distributional models make different predictions about the classification of a critical example midway between the nearest examples from each category. Participants' performance on such a critical example is evaluated in Experiments 1 and 2. This idea is extended in Experiments 3 and 4 where the effect of changing the relative variability of the two categories is investigated.

Modeling Sensitivity to Category Variability

To confirm the intuitive argument that exemplar and distributional models of categorization make opposite predictions two existing models of categorization, one where categorization is based on likelihood of belonging to representative distributions (general recognition theory, GRT, Ashby & Townsend, 1986), and the other where categorization is based on similarity to stored examples (the generalized context model, GCM, Nosofsky, 1986), are examined in this section. The ability of each to account for sensitivity to differences in the variability of approximately normally distributed categories is investigated.

Sensitivity of Exemplar and Distributional Models to Category Variability

The one dimensional category structure used in Experiments 1 and 2, is of the sort described in the intuitive argument above, where a critical test stimulus lies exactly half way between the nearest neighbors of two categories differing in variability. For each participant, for each category, eight examples were generated from a normally distributed category. The low variability category distribution had a standard deviation of 5.5 units, and the high variability category had a standard deviation of 14 units. There was a gap of 28 units between the nearest stimuli of each category, with the critical stimulus lying exactly in the center of this gap. To ensure

the gap between the nearest neighbors of each category was constant for all participants, for each participant the means of the categories needed to be adjusted slightly. Thus, for each participant, the low variability categories had the same population standard deviation, but slightly different means. The same is true of the high variability category for each participant. The result of this procedure is a set of categories with examples that are not evenly spaced, but are approximately normally distributed, and that meet the requirements that one category is more variable than the other, and that the critical stimulus lies between the nearest examples of each category.

The GCM and GRT will now be used to predict classification performance for the critical stimulus after training on the two categories. The modeling described here is for one participant's structure, although every structure used in the experiment was modeled and the results for each structure differed only very slightly, and did not alter the qualitative pattern of predictions for the GCM and GRT.

First consider the predictions of normal GRT. In normal GRT, the category examples are used to infer a population mean and variance for the normal distribution used to represent each category in perceptual space. A decision bound is then calculated that divides the perceptual space into regions for each category, so that all the stimuli represented by points in the same region are more likely to belong to the same category. The decision bound therefore corresponds to the line of stimuli that are equally likely to belong in either category. For two normally distributed categories the decision bound is a quadratic. In the one dimensional case the decision bound will be a single point. Perception is assumed to be noisy in GRT, and therefore each presentation of the same stimulus will not necessarily result in the

same percept. Thus a stimulus near the decision bound may sometimes be perceived to fall on one side of the bound, and sometimes on the other. To apply GRT to the category structure for Experiment 1 and 2 the eight examples for each category were used to generate an estimate of the population mean and a variance of the normal distribution for which the examples were drawn. (In fact, because perception is assumed to be noisy this method only provides the best estimate of a participant's hypothesized decision bound.) Then the point of equally likely classification was calculated, giving the decision bound (which in one dimension is a single point). The critical example was found to always fall on the high variability side of the decision bound, corresponding with the fact that this point was always more likely to have been generated by the high variability distribution. (Although the mean of the high variability category is further from the critical stimulus than the mean of the low variability category, the difference in variability between the two categories is sufficiently large that the critical stimulus is classified into the high variability category more often than the low variability category.) The exact predictions for classification of examples near the decision bound depend on the level of perceptual noise. Figure 5 illustrates this, showing the probability of classifying a stimulus in the high variability category as a function of the stimulus's value on the dimension. This function will be referred to as the generalization gradient. Throughout this paper the term generalization gradient will be the probability of classification of a stimulus as a member of the high variability category (and not the low variability category) as a function of the location of the stimulus. The exact shape of the generalization gradient depends of the perceptual noise associated with each stimulus. Figure 5 shows three gradients for three levels of perceptual noise. (After Ashby (1986) the noise is assumed to be Gaussian, with spherical covariance

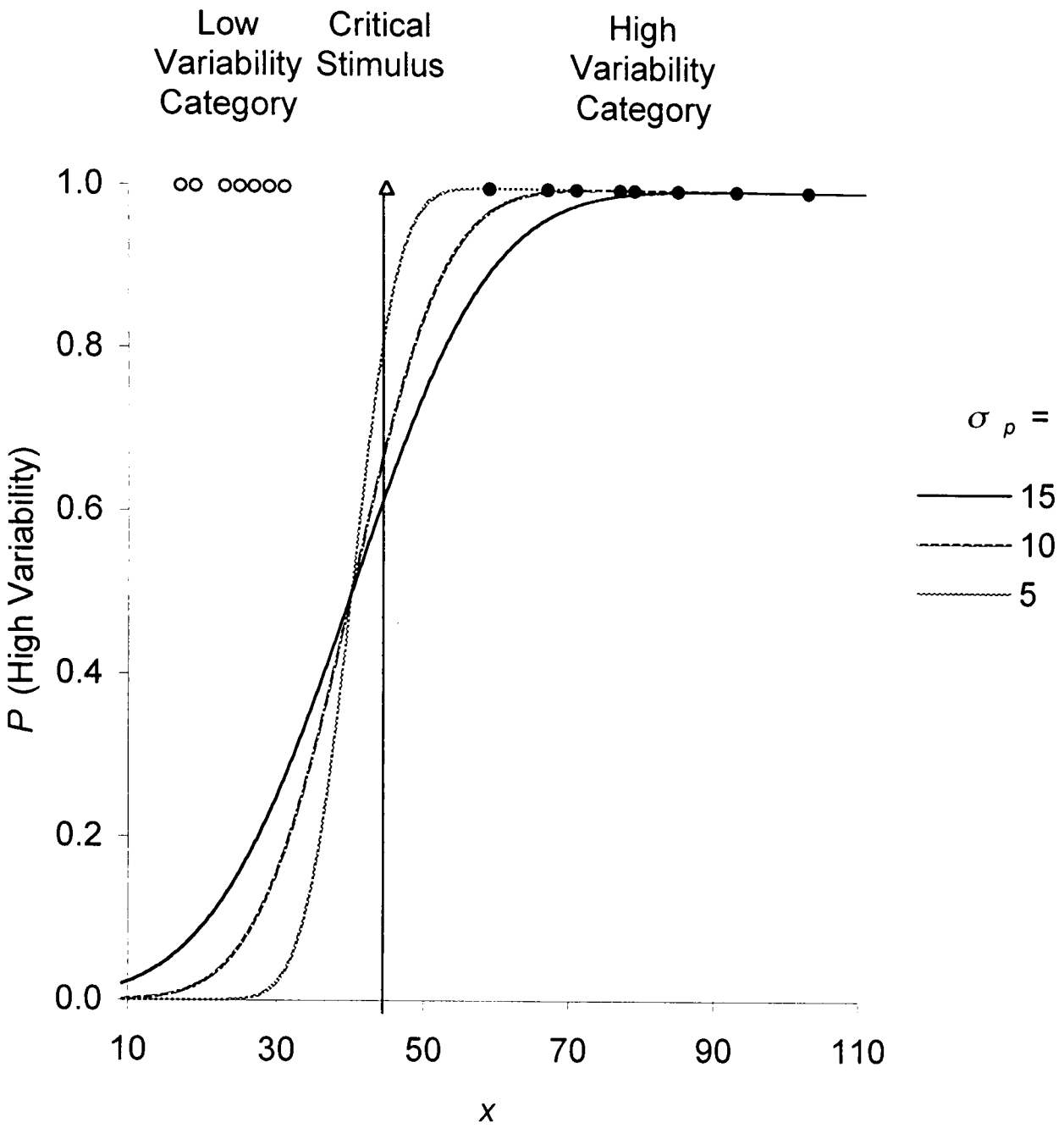


Figure 5. The probability of a high variability category response plotted as a function of the position of the stimulus on the dimension for normal GRT. The category structure is illustrated along the top of the figure, with one category more variable than the other. The three lines correspond to different levels of perceptual noise, which is assumed to be normally distributed with standard deviation σ_p.

matrix.) The less noise, the steeper the generalization gradient. In the case of perfect noiseless perception the GRT would have an infinitely steep generalization gradient (mathematically, the Gaussian would collapse into a delta function). In other words, the model would be completely deterministic, with the same stimulus always being classified the same way. Crucially though, the level of perceptual noise changes the slope of the generalization gradient, but does not alter the location of the decision bound. The perceptual noise never biases the decision bound one way or the other, and thus the critical stimulus is always more likely to be classified as a member of the high variability category.

Application of the GCM is more straightforward. To classify a new stimulus, the similarity between the new stimulus and each stored training stimulus is calculated. Luce's (1959) choice rule is then used to calculate probability that the stimulus is classified into either category from the stimulus's summed similarity to each category. Figure 6 presents generalization gradients, giving the probability that the stimulus is classified into the high variability category as a function of the stimulus's value on the dimension. The different gradients correspond to predictions of the GCM with different values of the generalization parameter, c. As the amount of generalization increases, examples more distant from the to be classified stimulus are weighted more heavily. Thus for small values of the generalization parameter, when generalization is high, the generalization gradient is steeper. When the amount of generalization is large, the critical stimulus is always predicted to be classified into the low variability category. As the amount of generalization is reduced (and the GCM approximately tends towards a nearest neighbor model), the critical stimulus becomes less likely to be classified in the low variability category. Crucially though, the model is never able to predict that the critical stimulus is more likely to be

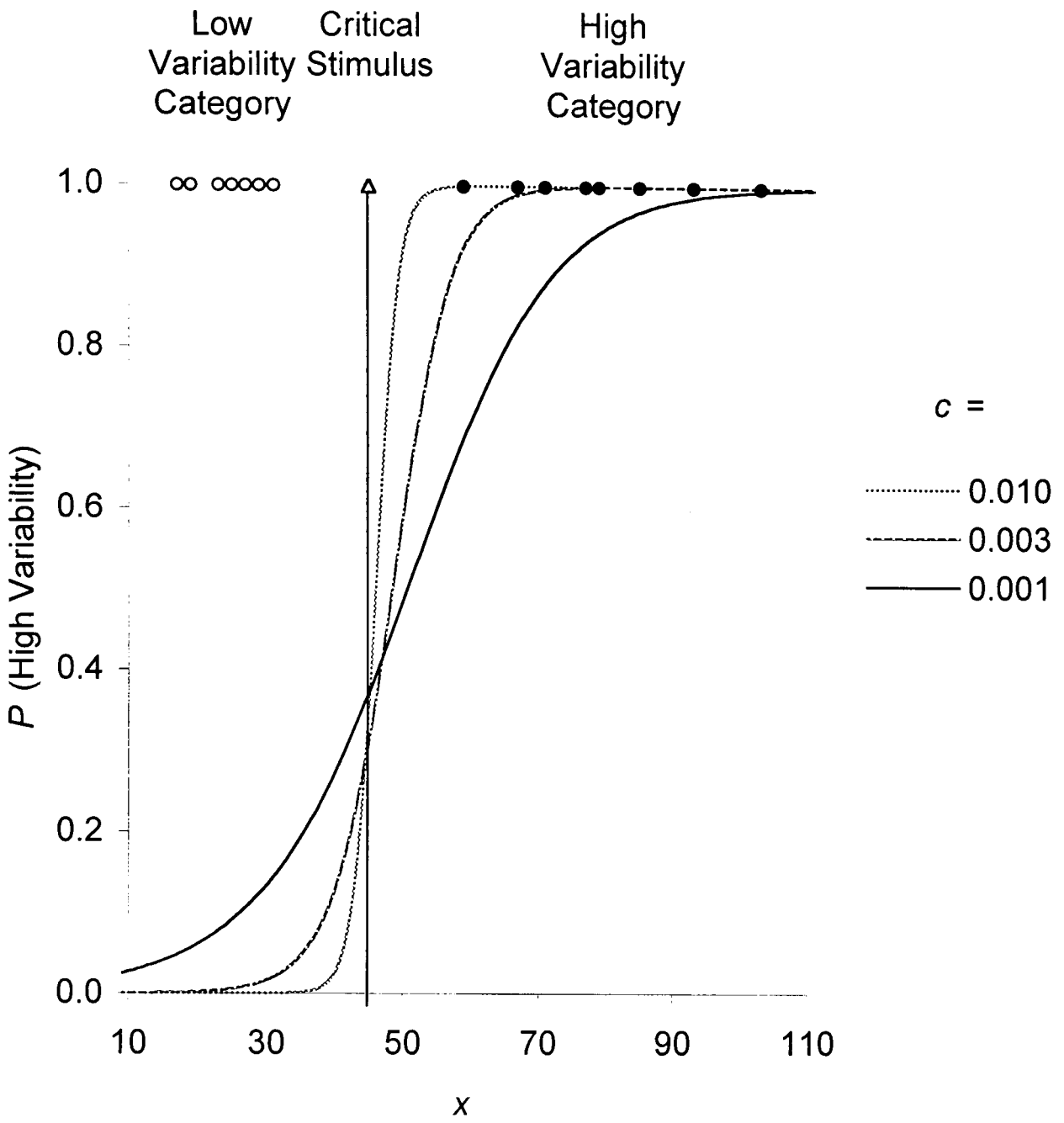


Figure 6. The probability of a high variability category response plotted as a function of the position of the stimulus on the dimension for GCM (q=2). The category structure is illustrated along the top of the figure, with one category more variable than the other. The three lines correspond to different values of the generalization parameter, c.

classified into the high variability category. The predictions here are for the GCM with a Gaussian function relating similarity to distance (q=2). The predictions of the GCM with an exponential similarity function (q=1), which is used when stimuli considered easily discriminable, do not differ qualitatively from the Gaussian similarity function GCM.

In summary, for a critical stimulus that lies exactly between the nearest neighbors of two categories that differ in variability, provided the difference is variability is sufficiently great, the GCM predicts the critical stimulus is more likely to be classified into the low variability category (independent of the amount of generalization), and GRT predicts the critical stimulus more likely to be classified into the high variability category (independent of the amount of perceptual noise).

Sensitivity of Exemplar and Distributional Models to Changes in the Relative Variability of Categories

Experiment 3 uses two new two dimensional pairs of categories to investigate how changing the relative variability of two categories should affect the classification intermediate stimuli, according to the GCM and GRT. Both category structures have two categories, one with a mean of (200, 200) and the other with a mean of (300, 300). In fact 10 examples of each category are arranged in a circle around each mean (Figures 7 and 8). The diameter of one circle is larger than the other, such that one category is more variable than the other. In the 1:2 pair of categories the low variability category is half as variable than the high variability category (standard deviation of 20 vs. 40), and in the 1:4 pair of categories the low variability structure is four times less variable than the high variability category (standard deviation of 12.7 vs. 50.2). The transfer examples are used to measure participants' generalization between the two categories in Experiments 3 and 4.

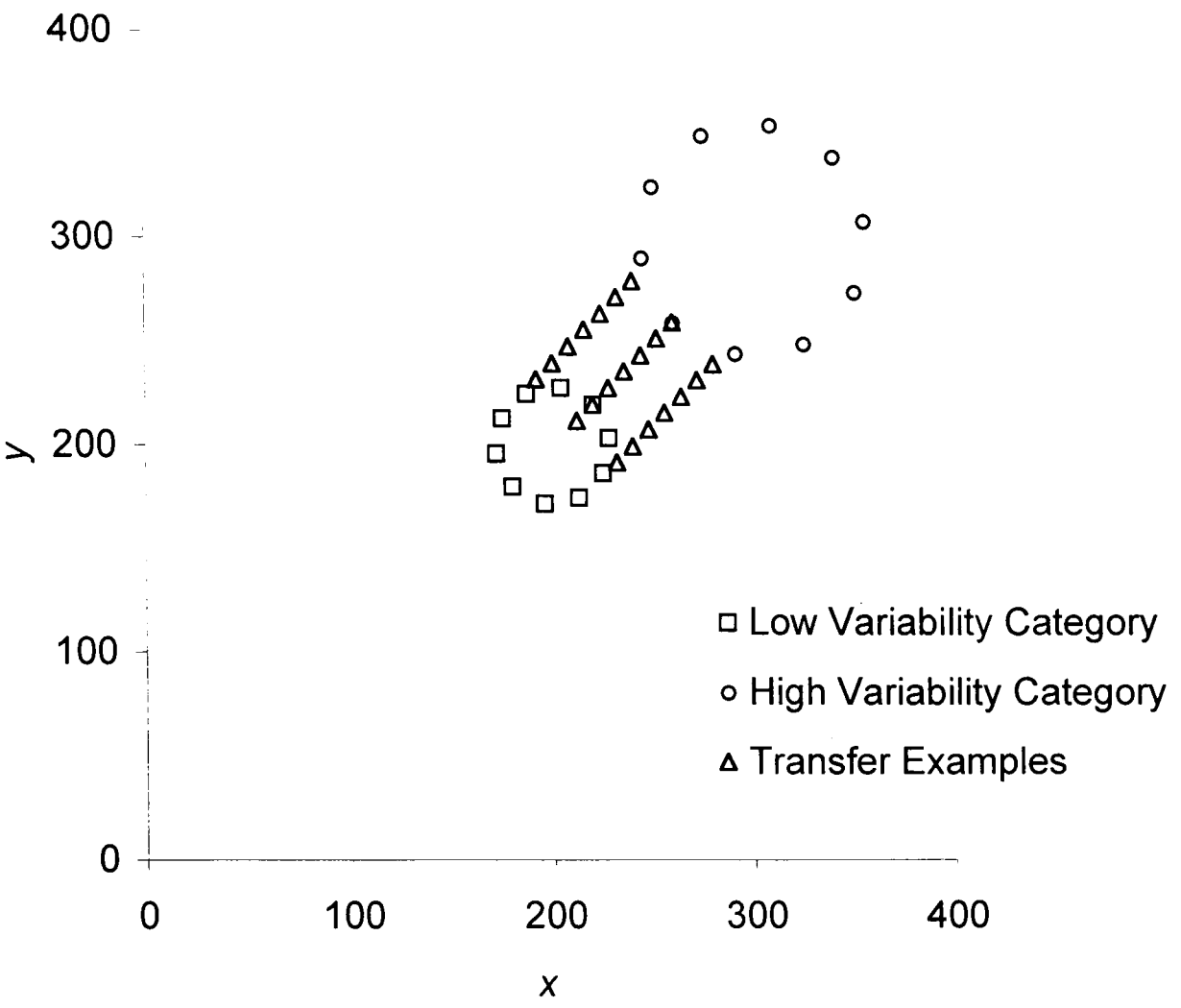


Figure 7. The arrangement of examples for the 1:2 pair of categories.

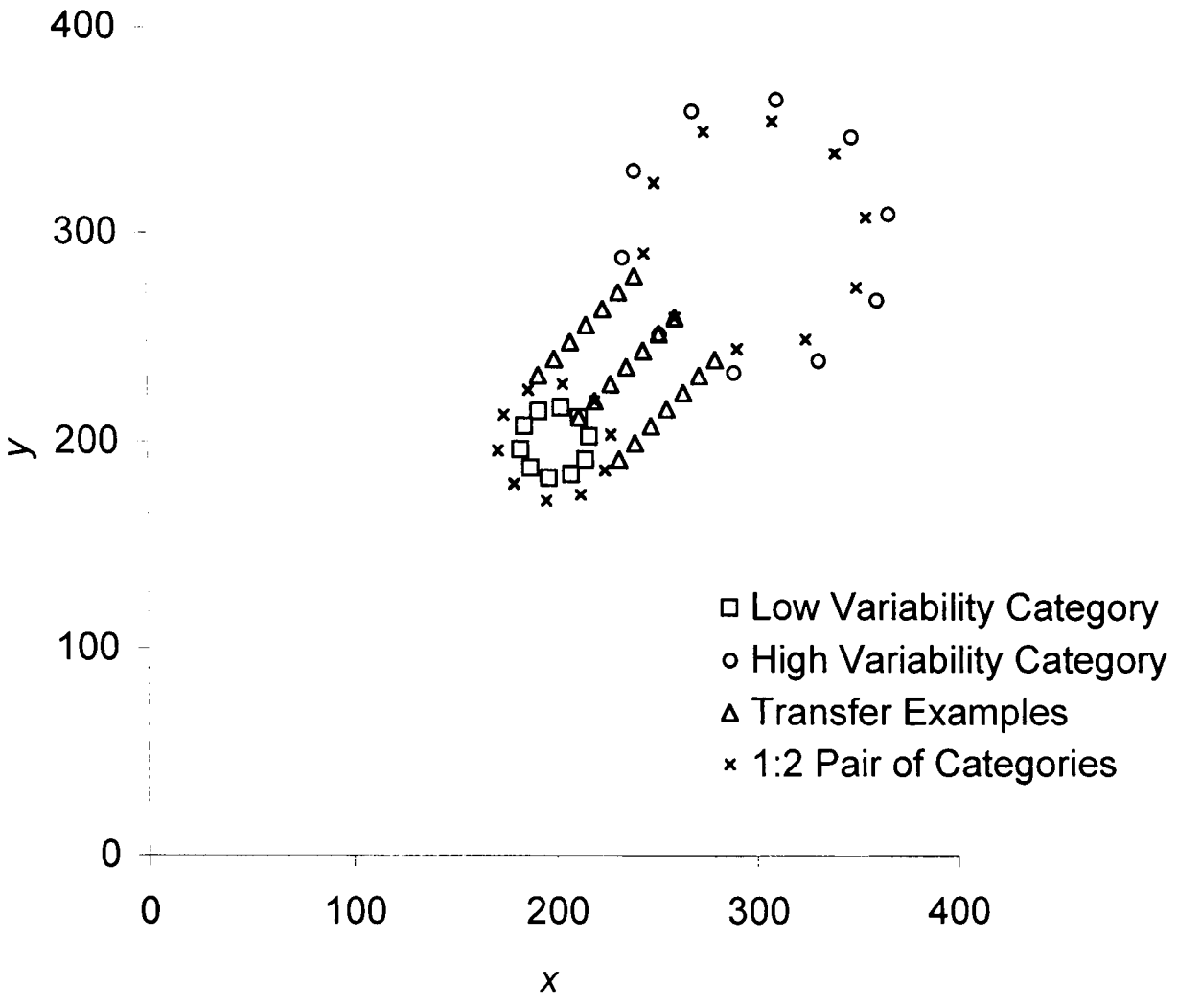


Figure 8. The arrangement of examples for the 1:4 pair of categories. The training examples from the 1:2 pair of categories are shown for comparison.

Given the category representation of the normal GRT, it seems likely that this model will be sensitive to differences in the relative variability of two categories. This is indeed the case. All the categories are represented using simple co-variance matrices ($\underset{\sim}{\Sigma}=\underline{\sigma}^2\underset{\sim}{I}$) due to the symmetrical nature of the categories. For the 1:2 pair, three generalization gradients (the probability of a high variability response for stimuli on the line $\underline{y}=\underline{x}$ as a function of the stimulus's $\underline{x}$ value) were calculated for the different levels of perceptual noise, and are shown in Figure 9. As in modeling for the one dimensional category structure used in Experiments 1 and 2, the perceptual noise changes the slope of the generalization gradient, but does not bias the decision bound one way or the other. Of interest here is the comparison of gradients for the 1:2 and 1:4 pairs. The one generalization gradient for each category structure is shown in Figure 10. (The level of perceptual noise is assumed constant across both structures, $\underline{\sigma}_p=10$.) As the difference in variability between the two categories is increased the decision bound moves nearer the low variability category.

The particular values of variability for each category were chosen to keep the distance between the nearest two examples of each category constant across the 1:2 and 1:4 pair. This allows comparison of the classification of items that are either the same distance from the means of each category (i.e., with the same absolute co-ordinates), or comparison of the classification of items that are the same distance either from the nearest neighbor of the low variability category, or the nearest neighbor of the high variability category (i.e., with the same co-ordinates, relative to the nearest neighbors). For comparison of items with either the same absolute co-ordinates (Figure 7), or the same co-ordinates relative to the nearest neighbors (Figure 11), the item is always more likely to be classified into the high variability category in the 1:4 pair compared to the 1:2 pair. This is always true, provided the

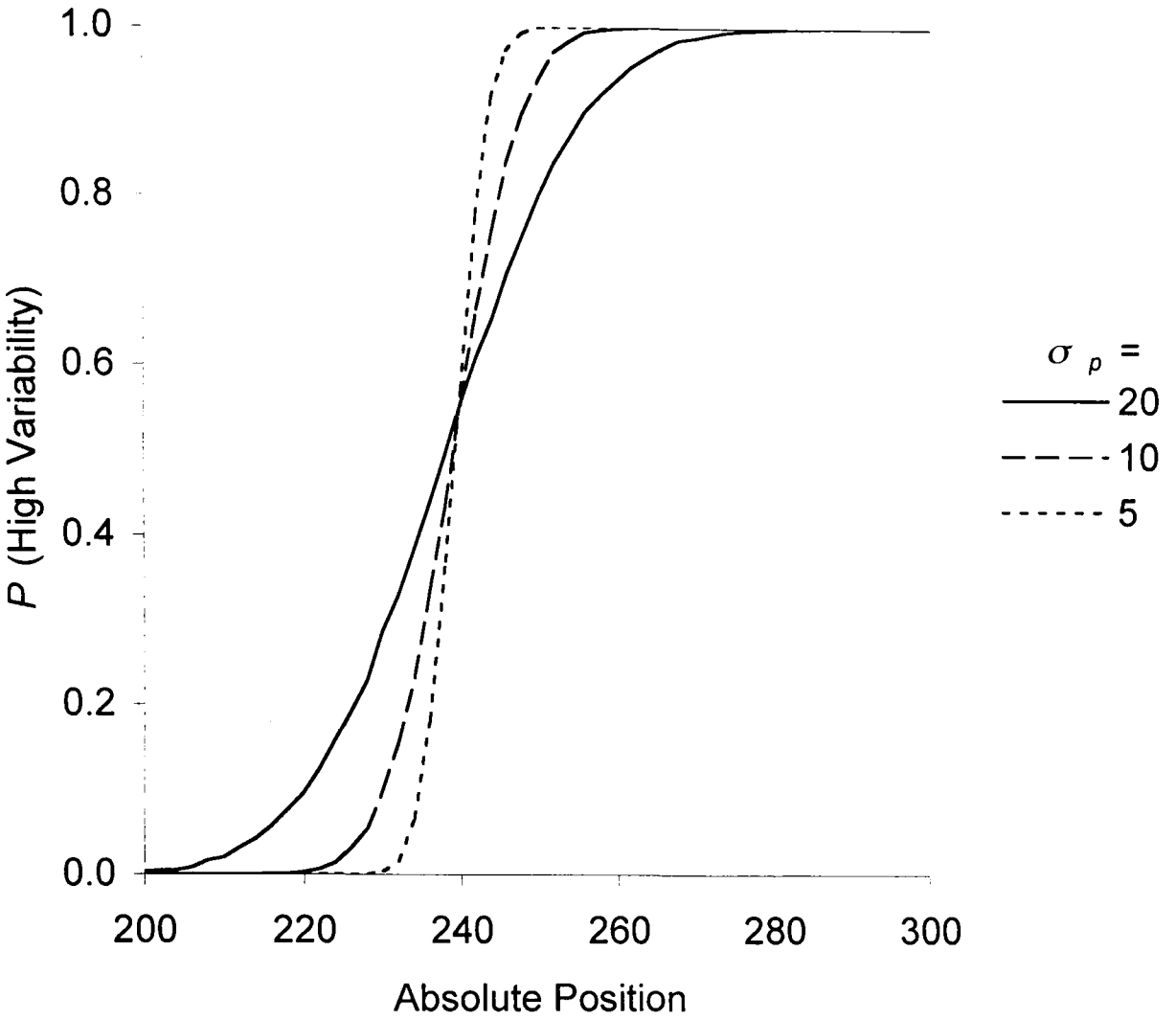


Figure 9. The probability of a high variability category response for an example on the line y=x plotted against the x value of the absolute position of the example for normal GRT for two categories (pair 1:2), one twice as variable than the other.

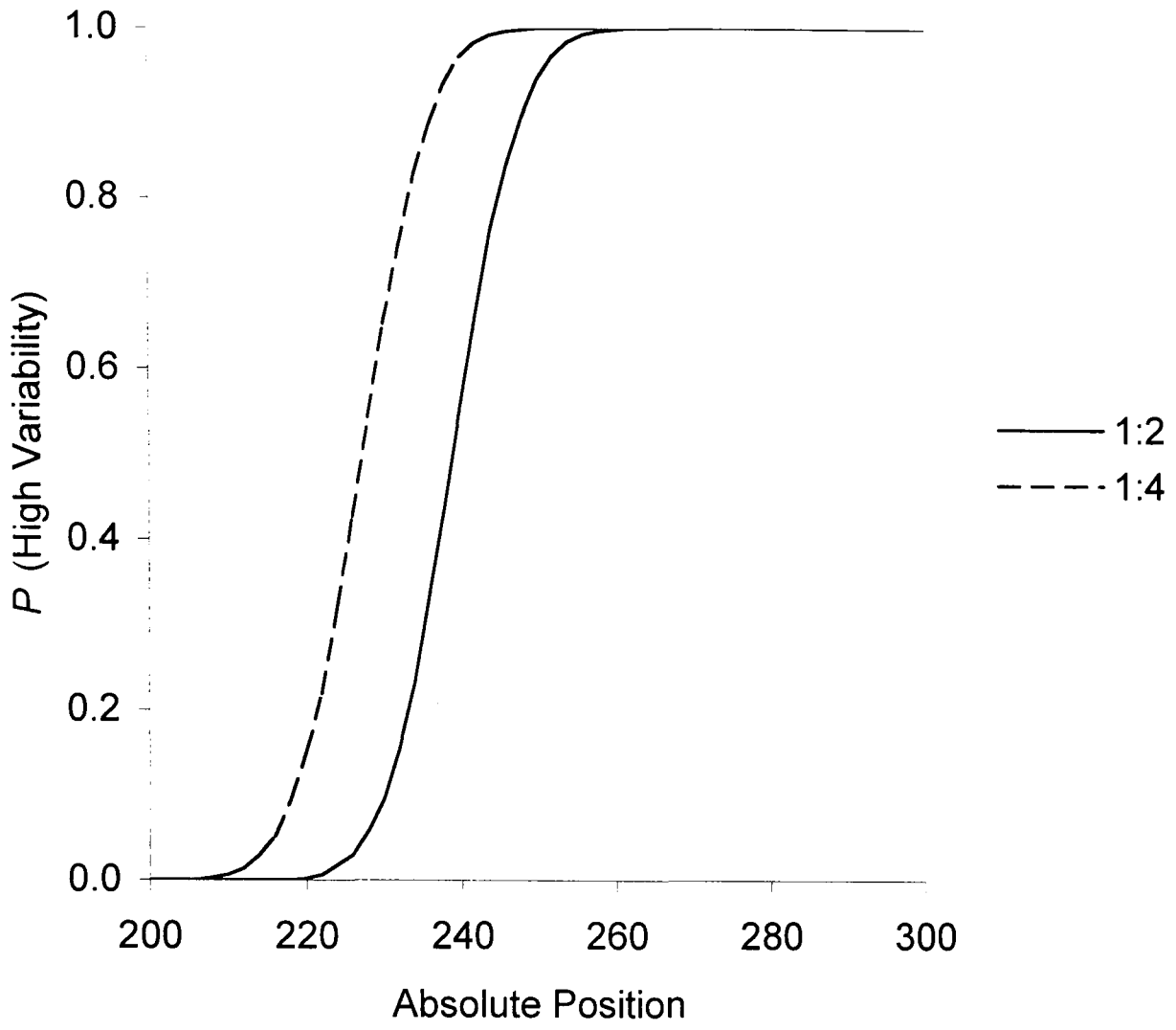


Figure 10. The probability of a high variability category response for an example on the line y=x plotted against the x value of the absolute position of the example for normal GRT (σ_p=10) for two pairs of categories, pair 1:2 and pair 1:4.

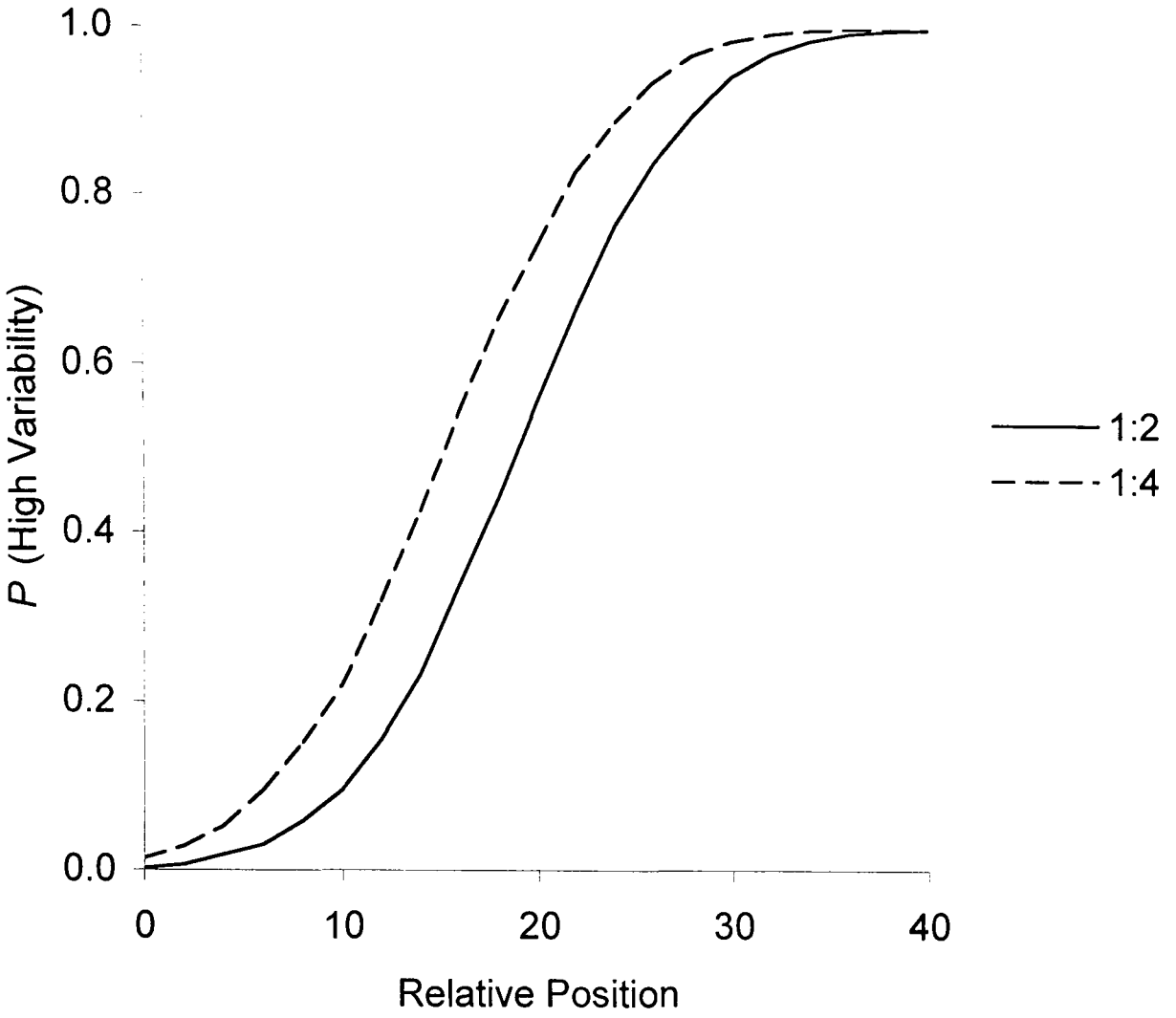


Figure 11. The probability of a high variability category response for an example on the line $y=x$ plotted against the x value of the example relative to the nearest neighbor of the low variability category for normal GRT ($\sigma_p=10$) for the 1:2 and 1:4 pairs of categories. This figure is equivalent to Figure 10 with the 1:4 pair line shifted 12 units closer to the 1:2 pair line.

difference in the relative variability of the two categories is great enough across the two category pairs.

In summary, the slope of the generalization gradient is given by the amount of perceptual noise, and the location (the points of equal classification into either category) is determined by the relative variability of the two categories – the greater the difference, the closer the decision bound is to the low variability category.

The generalization gradients predicted by the GCM for the two category structures are shown in Figure 12, with the generalization parameter, c, held constant across the two structures. The predictions here are for the GCM with a Euclidean distance metric and a Gaussian similarity function (q=2, r=2), however the pattern of the predications is the same for a city block distance metric and exponential similarity function (q=1, r=1). Traditionally, the city block exponential GCM is used to model categorizations where the stimulus dimensions are considered separable, and the Euclidean Gaussian GCM is used when dimensions are integral (Garner, 1974; Nosofsky, 1987). The predictions of the GCM are very similar to those of GRT, with intermediate items being more likely to be classified as members of the high variability in the 1:4 pair than the 1:2 pair. This is because in the 1:4 pair of categories, the high variability category's items are nearer, and the low variability category's items are further away from a given intermediate item, compared to the 1:2 pair. However, when the generalization gradients are measured relative to the two nearest neighbors, rather than relative to the category means, this is no longer true (Figure 13). Now in the two category structures, for a given intermediate item, the nearest neighbor items of each category are equally distant. However, in the 1:4 pair, because the low variability category items are less spread out, the second nearest item in the low variability category will be nearer than the corresponding

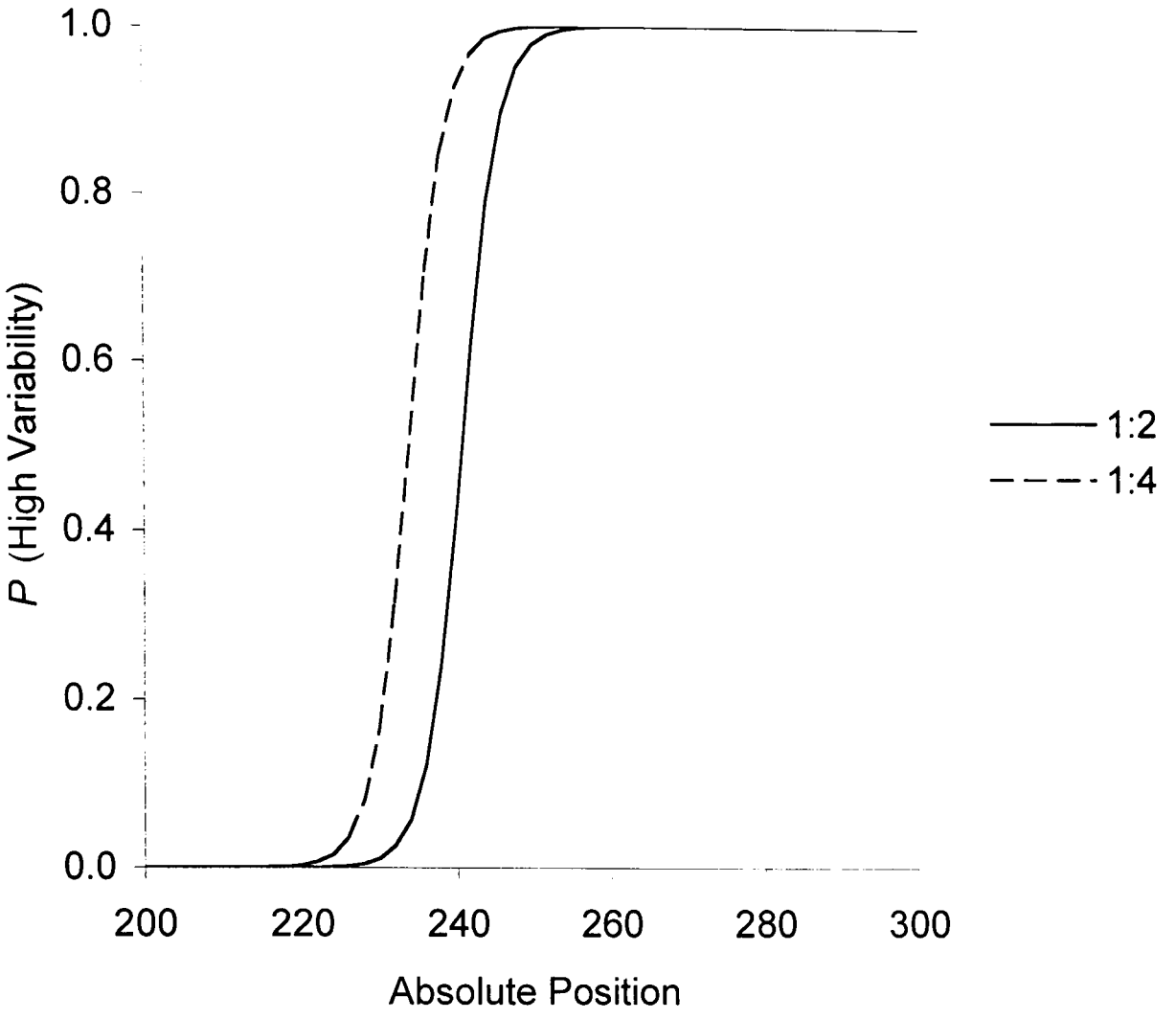


Figure 12. The probability of a high variability category response for an example on the line y=x plotted against the x value of the absolute position of the example for the GCM (q=2, r=2, c=0.05) for two category pairs 1:2 and 1:4.

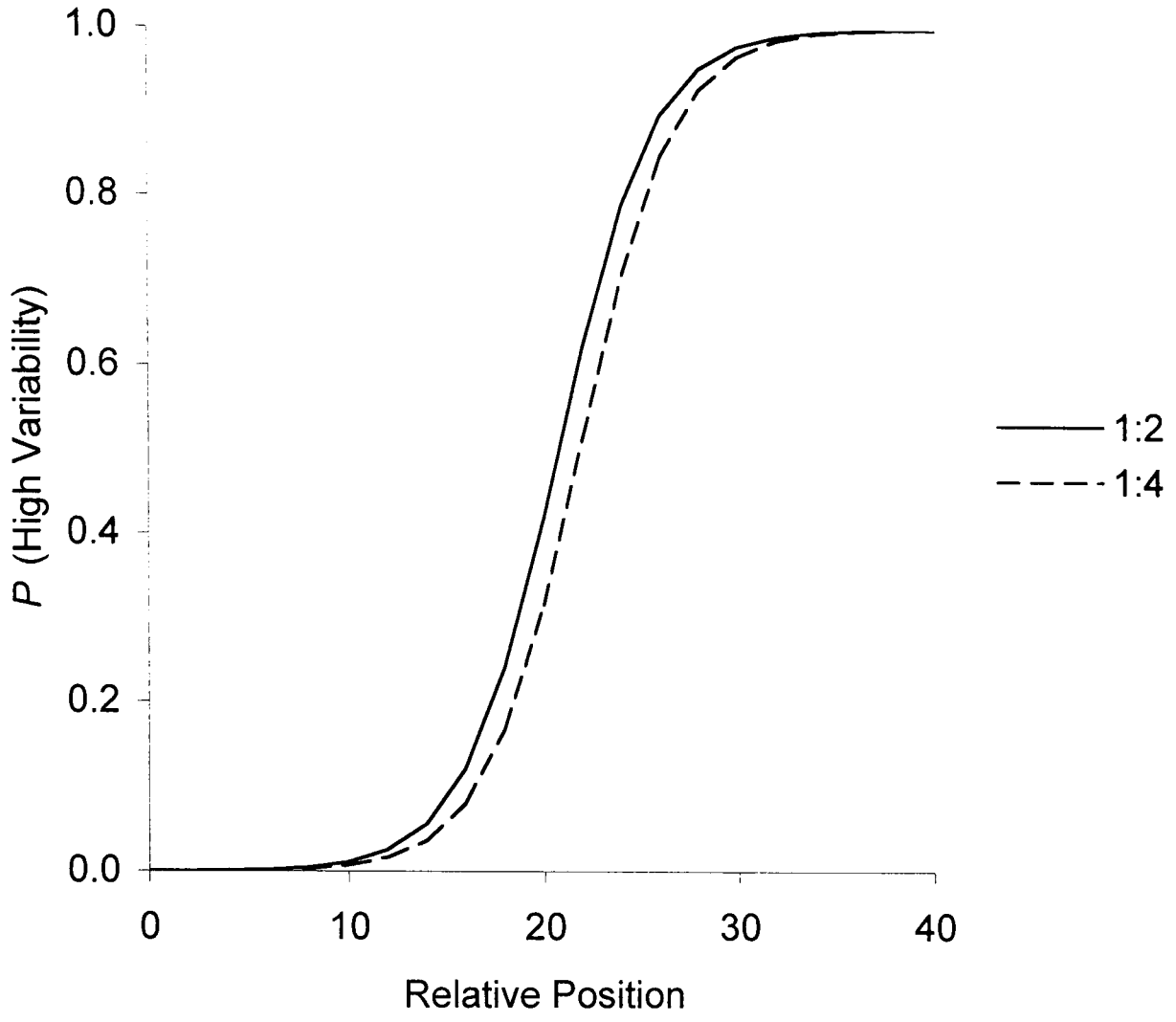


Figure 13. The probability of a high variability category response for an example on the line y=x plotted against the x value of the the example relative the the nearest neighbour of the low variability category for the GCM (q=2, r=2, c=0.05) for category pairs 1:2 and 1:4. This figure is equivalent to Figure 12 with the 1:4 pair line shifted 12 units to the right.

item in the low variability category of the 1:2 pair. Similarly, in the 1:4 pair, because the high variability category items are more spread out, the second nearest item in the high variability category will be further away than the corresponding item in the high variability category of the 1:2 pair. Thus, when items equally distant from the nearest examples in the two category structures are compared, the item is more likely to be classified as a member of the high variability category in the 1:2 pair compared to the 1:4 pair – the opposite prediction to GRT.

Experiment 4 uses the 1:2 pair described above and a new pair of categories. This new pair, 1:2 expanded, differs only slightly from the 1:2 pair – in the 1:2 expanded pair the 5 items of the high variability category that are furthest from the low variability category are moved to even more extreme points (Figure 14). This pair is designed to allow the exemplar model to be further tested. Figure 15 shows the generalization gradients predicted by the GCM (q=2, r=2, c=0.05) for the two conditions. The gradients almost exactly coincide. This is true for the range of c parameters that produces acceptable accuracy for the training examples (i.e., greater than 80% accuracy – participants in fact performed at about 90% accuracy). This can be intuitively explained as follows. When the amount of generalization is small (i.e., c is large), old training items will be will be accurately classified, as their classification is determined mainly by the category label stored for the exemplar representing the old training item. As the amount of generalization is increased, other exemplars have more and more influence on the classification of the old training item. If the amount of generalization is too high (i.e., c is too small), then exemplars from the other category will influence the classification of the training item, reducing the predicted accuracy with which the item will be classified. When classifying items from one category the amount of generalization must be small enough to prevent

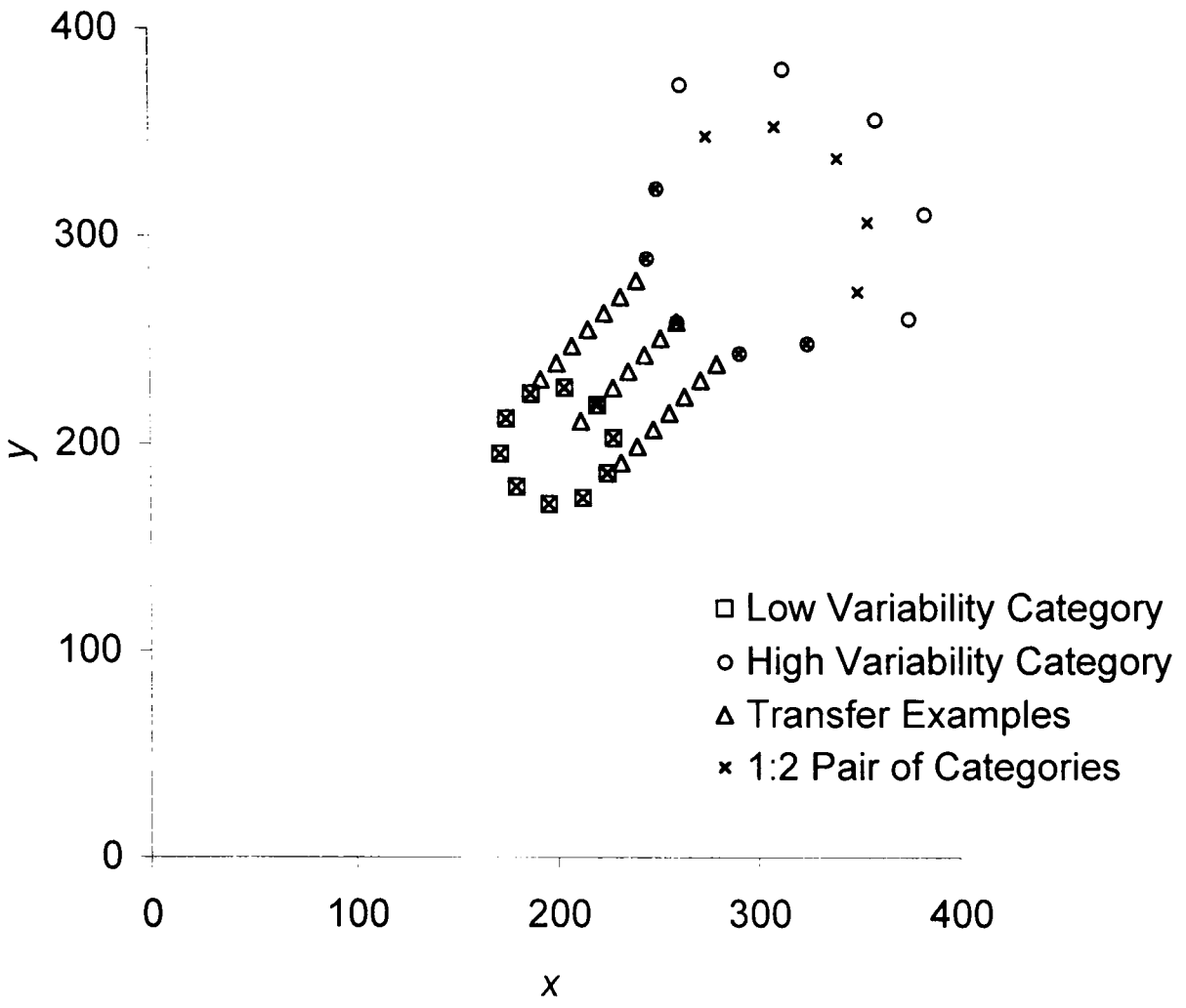


Figure 14. The arrangement of examples for the 1:2 expanded pair of categories. The training examples from the 1:2 pair of categories are shown for comparison.

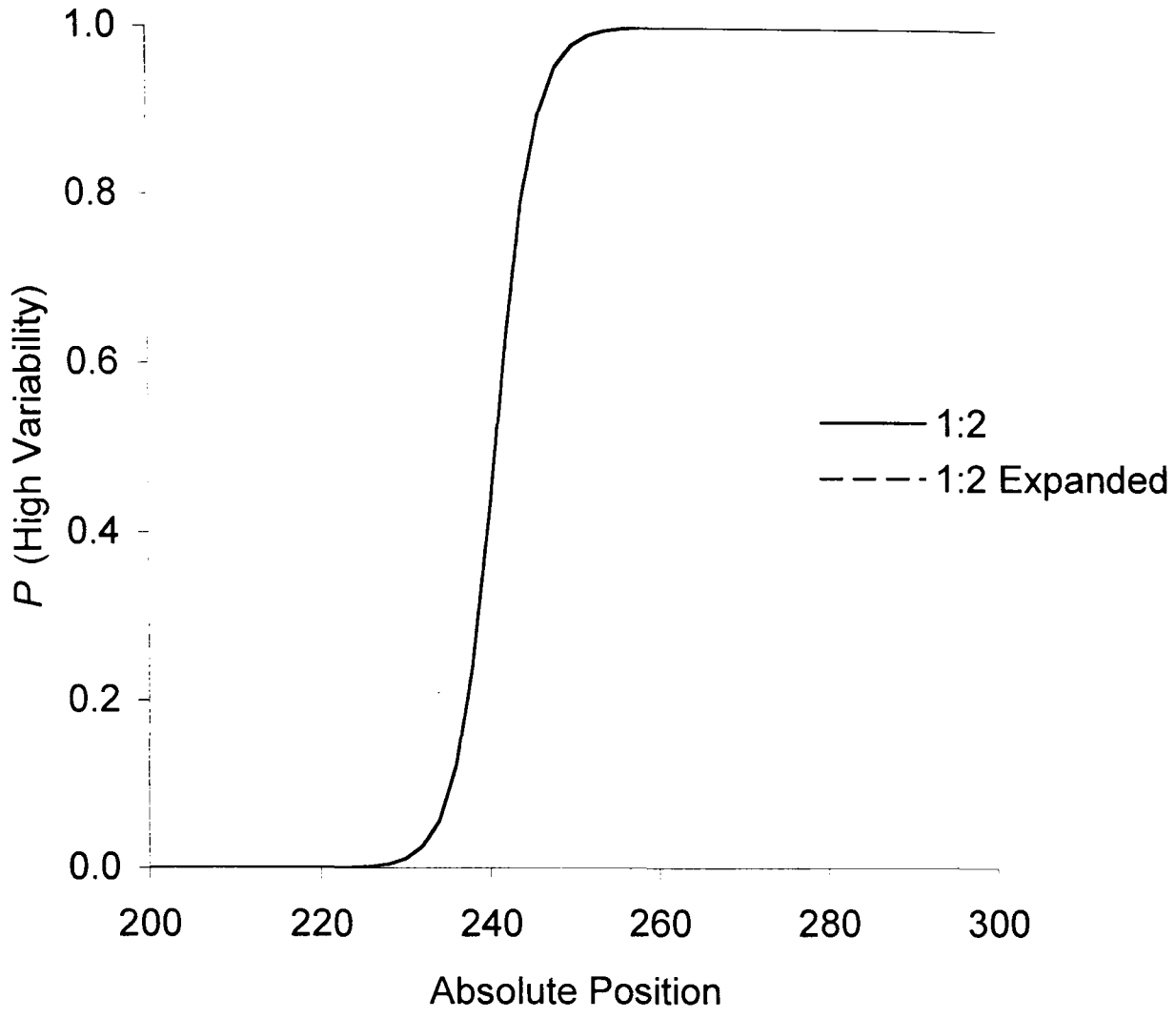


Figure 15. The probability of a high variability category response for an example on the line y=x plotted against the x value of the absolute position of the example for the GCM (q=2, r=2, c=0.05) for two pairs of categories 1:2 and 1:2 expanded. Notice that the two gradients exactly coincide.

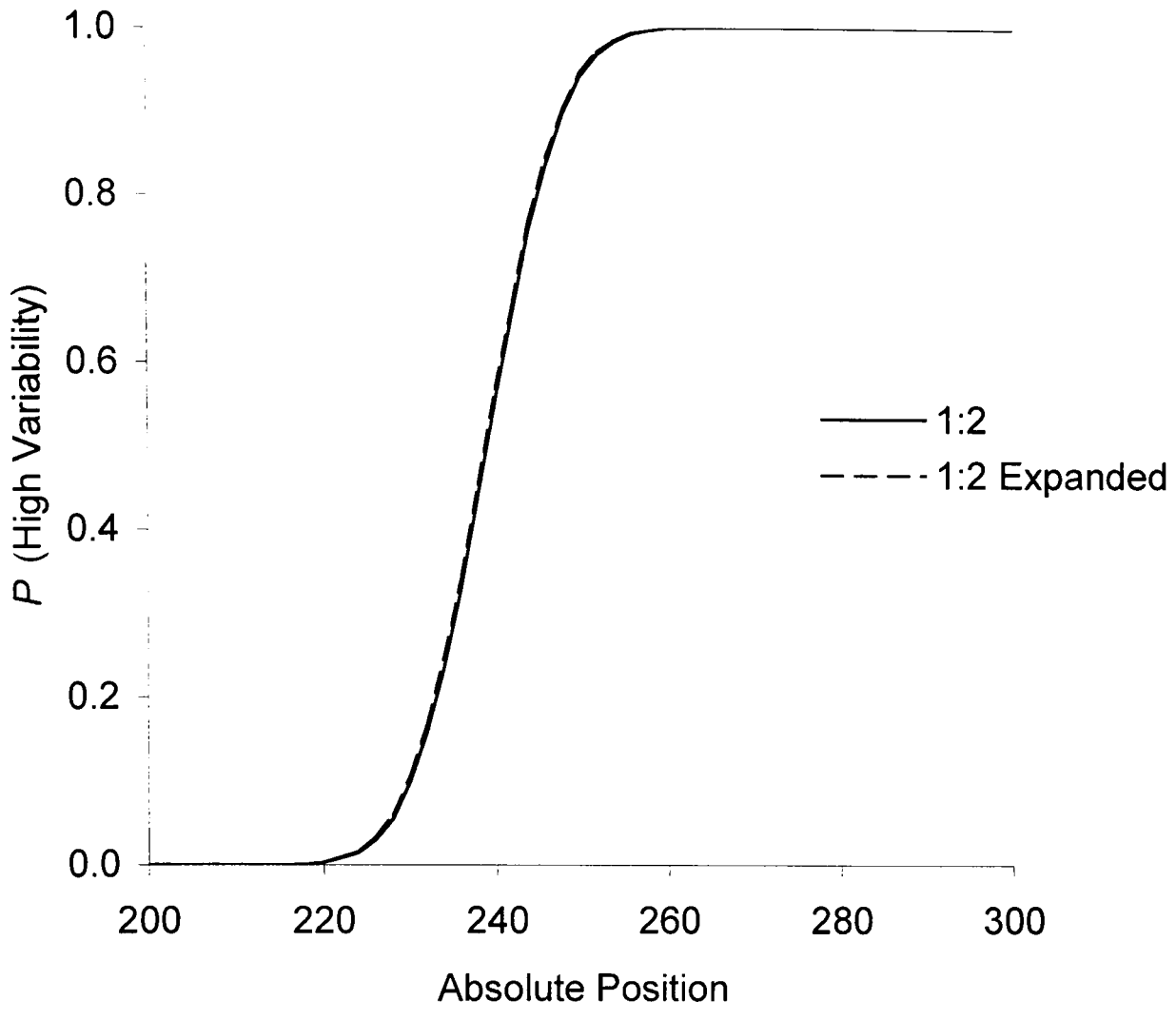


Figure 16. The probability of a high variability category response for an example on the line $y=x$ plotted against the x value of the absolute position of the example for the GRT ($\sigma_p=10$) for two pairs of categories 1:2 and 1:2 expanded. Notice that the two gradients exactly coincide.

Experiment 1

Experiments 1 and 2 were designed to discriminate between exemplar-based classification and distribution-based classification using a category structure as described above. Some participants in the experiment were given a hint telling them that the two categories differed in variability. Smith and Sloman's (1994) replications of Rips' (1989) study suggest that participants only categorize stimuli into the high variability category when their verbal protocols show awareness of a difference in variability between the two categories. The hint here was included to see what effect knowledge of the variability might have on participants' classification of these perceptual stimuli. The stimuli used in this experiment are circles with a dot on the circumference, where the position of the dot on the circumference varies between stimuli. Pilot studies used the position of the dot on a straight line, but the performance of many participants was consistent with their reports of using a rule, such as whether the dot was more or less than half way along the line, to make their decision. The stimuli here were chosen so that use of rules like this should not be possible.

Method

Participants. 32 undergraduate students from the University of Warwick participated for course credit. Participants were alternately assigned to one of two conditions, labeled the "hint" or "no hint" conditions.

Design. Participants successively performed three binary categorization tasks, between high and low variability categories defined on a single underlying stimulus dimension. After categorizing 16 'training' examples, participants classified a critical stimulus that fell half way between the nearest example of the low variability category and the nearest example of the high variability category. They

then classified two further examples, one from each category, before moving on to the next classification. Half of the participants were given a hint, described below, during the instructions at the beginning of the experiment pointing out that one category was more variable than the other.

Stimuli. The stimuli used in this experiment were black outline circles each with a single solid black dot, somewhere on its circumference. The stimulus was presented on a white background. The diameter of the circle subtended approximately 2 degrees of visual angle. The dot was $1/6^{th}$ the diameter of the circle. The stimuli varied only in the position of the dot around the circumference and this position was diagnostic of category membership. Three stimulus sets were constructed for each participant. A stimulus set met the following criteria. There were eight examples for each of the two categories. The examples of one category, the low category, had a standard deviation of 5.5 degrees variation in the position of the dot on the circumference. The examples of the other category, the high category, had a standard deviation of 14 degrees. The two categories did not overlap, and always had a fixed distance of 28 degrees between the nearest examples of the two categories. The critical stimulus was always exactly half way between the nearest example of the low category and the nearest example of the high category. The position of the critical stimulus in each participant's three stimulus sets varied. The critical stimulus was in the 45 degree position for the first task, the 135 degree position for the second task, and the 225 degree position for the third task (with 0 degrees being at the 12 o'clock position, and angle increasing clockwise). This ensured that a horizontal, vertical, or diagonal diameter could not be used to classify the critical stimulus. The relative mean position of the low and high variability categories was counterbalanced across participants. That is, for half the participants

the category clockwise of the critical stimulus was the low variability category, and for the other half the category was the high variability category.

For each participant the stimulus set was modeled to check that the critical stimulus was indeed more similar to the low variability category, but more likely to belong to the high variability category.

Apparatus. Stimuli were displayed on a 14" Apple Macintosh Color Display. Responses were collected using labeled keys on a standard qwerty keyboard. The keys A to J inclusive were labeled A, B, C, yes, D, E, F from left to right.

Procedure. Participants were alternately assigned to either the hint or no hint condition. They were seated comfortably in front of the computer, and the keyboard and monitor were adjusted as necessary. The experiment began with instructions displayed on the computer screen. Participants were told they would do three categorization tasks, one after the other. They were told that they would discriminate types A and F, then types B and E and finally types C and D. The structure of a trial was described, and participants were asked to respond as quickly as they could without making mistakes. Participants in the hint condition received further instructions:

> "IMPORTANT HINT: For each pair of types that you learn about, one type is allowed a greater spread of dots than the other. During the experiment try to work out which type this is. For example, when learning about types A and F, try to decide whether it is type A or type F that is allowed a greater spread of dots. Do not forget your hint. It is very important."

When the experimenter had checked that a participant understood the instructions, and if applicable, that they understood the hint, the experiment began with the first trial. A ready prompt appeared on the screen. When a participant pressed yes, there was a 1.5 s blank screen before a circle with a dot appeared on the screen for 1 s. The participant responded as quickly as they could from stimulus

onset. The assignment of category labels to the high and low variability categories was counterbalanced across participants. After one second the screen was cleared, whether the participant had responded or not. After the participant responded the correct answer was displayed on the screen for 1.5 s, followed by a 1.5 s blank screen before the next trial began. The first 19 trials used examples from the first stimulus set, the second 19 the second set and the third 19 the third set. Within each block of 19 trials the first 16 were the 8 examples from each of the two categories, in a random order. The 17th trial was the critical stimulus. The feedback given on this trial was random. The final two trials were re-presentations of an example from the center of each category, and were used to provide an accuracy measure that allowed performance in this experiment to be compared to performance in Experiment 2. The order of stimuli on trials 18 and 19 was counterbalanced across participants. After the 19th trial on a given stimulus set participants moved on to the next set.

Results

Table 1 displays the mean proportion of training trials correct, the mean proportion of verification trials correct and the proportion of high variability category responses for the critical stimuli. Average training accuracy was consistently high, despite the small number of training trials. There was no effect of hint on the mean proportion of training trials correct across all blocks, $t(31)=0.88$, $p>0.05$. Knowledge that the two categories differ in variability did not facilitate category learning. This was confirmed by performance on the two verification trials that followed each critical stimulus trial. Verification accuracy was high and was not affected by hint, $t(31)=0.85$, $p>0.05$.

Of most interest was the result that in both conditions the mean proportion of high variability responses was significantly below chance level of 0.5: $t(15)=13.17$,

Table 1

Means across all three blocks for Experiment 1. (Numbers in brackets are standard errors of the means.)

	Mean training proportion correct	Mean verification proportion correct	Mean proportion of high variability category responses to critical stimuli
Hint	0.79 (0.02)	0.90 (0.04)	0.38 (0.09)
No hint	0.81 (0.02)	0.94 (0.03)	0.25 (0.06)

$p<0.05$ for the no hint condition; $t(15)=7.31$, $p<0.05$ for the hint condition. Participants favored categorizing the critical stimulus into the low variability category. The mean proportion of high variability responses to critical stimuli does not differ significantly between the hint / no hint conditions, $t(31)=1.22$, $p>0.05$. Participants were not significantly influenced by explicit instructions informing them the two categories differed in variability. The numerical difference between the conditions is however in the direction expected if the hint increases participants' sensitivity to categorical variability. Individual participants were not always consistent in their responding to the critical stimulus (i.e., always responding with the low variability category, or always responding with the high variability category), although as there were only three trials for each participant there is not enough data to further explore this hypothesis.

Discussion

In this experiment a stimulus lying midway between the nearest examples of two categories differing in their variability was more likely to be classified as belonging to the lower variability category, in line with the prediction of the exemplar view. Telling participants that one category was more variable than the other did not have a significant effect on their performance. This raises the question of whether highlighting the difference in variability more strongly would encourage participants to change their classification strategy. This was the motivation for Experiment 2.

Experiment 2

Experiment 2 was designed to make the difference in variability of the two categories more salient. Experiment 2 uses the same stimulus set as Experiment 1. The primary difference between the experiments is that in this experiment the entire

set of training examples are presented simultaneously.

Method

Participants. 32 undergraduate students from the University of Warwick participated for course credit. Participants were alternately assigned to either the hint or no hint conditions. Participants had not taken part in Experiment 1.

Stimuli. The stimulus structure used in this experiment is identical to the structure used in Experiment 1. In fact, the first participant in this experiment saw the same stimuli as the first participant in Experiment 1, and so on. The only difference is that now stimuli are presented simultaneously, on paper. Each circle had a diameter of 25 mm.

Procedure. Participants were given written instructions telling them they would learn three categorization tasks, one after the other. They were told that for each categorization task they would see a sheet of examples that they should study for one minute. They were told that they should pay attention to the position of the dot, and the category labels, because this information is what would help them classify three test examples presented afterwards. Participants in the hint condition were given the further instructions, informing them that the two categories would differ in variability, and they should try to work out which category was the more variable for each categorization.

After the experimenter had checked that the participant had understood the instructions the examples of the first stimulus set were displayed. The 16 examples were arranged on the same piece of paper. Each set of eight examples belonging to the same category was arranged in a row, inside a rectangle, together with the category label. The two sets were placed one above the other. The placement of the low and high variability categories at the top and bottom of the page was

counterbalanced across participants, as was the assignment of labels to categories. Within a set, for all participants, the examples were arranged in the same rank order. This was to ensure that if the order of the examples on the page affected the salience of the variability, then it would be held constant across the hint no hint conditions. Participants studied the sheet of examples for 1 minute, before it was removed from sight. The critical stimulus was then presented in the center of a new piece of paper, together with the instruction, "Is this circle with dot a bip or a bap. Please circle your answer." Participants circled the category label they felt the example belonged to. This was repeated with two old examples, one from the center of each category. This was to check that participants had some memory of the categories and corresponding labels, and had not simply guessed. Participants then moved onto the next stimulus set.

Results

The analysis of results here is similar to the analysis for Experiment 1. The only difference is that because results for training accuracy could not be gathered, they cannot be analyzed here.

Table 2 displays the mean proportion of verification trials correct and the proportion of high variability category responses for the critical stimuli. Verification accuracy was high and was not affected by hint, $t(31)=0.88$, $p>0.05$. The level of accuracy on the verification stimuli was similar to the accuracy on the same stimuli in Experiment 1.

The categorization of the critical stimulus is of greatest interest. The mean proportion of high variability responses to critical stimuli does differ between the hint/no hint conditions, $t(31)=2.23$, $p<0.05$. In the no hint condition performance on the critical stimuli is not different from chance, $t(15)=0.13$, $p>0.05$. In the hint

Table 2

Means across all three blocks for Experiment 2. (Numbers in brackets are standard errors of the means.)

	Mean verification proportion correct	Mean proportion of high variability category responses to critical stimuli
Hint	0.92 (0.04)	0.74 (0.07)
No hint	0.96 (0.02)	0.51 (0.08)

condition, the proportion of high variability responses was significantly above chance, $t(15)=3.61$, $p<0.05$. Participants favored categorizing the critical stimulus in the high variability category when given the variability hint. As in Experiment 1, participants were not always consistent in their responding.

Discussion

In the variability hint condition, participants classified a stimulus midway between the nearest examples of two categories differing in variability as belonging to the higher variability category significantly more often than chance. With no hint, their performance on this critical stimulus was almost exactly at chance. By comparing these results with the results from Experiment 1 it can be seen that increasing the salience of the difference in the variability of the two categories increased the proportion of times was classified as belonging to the high variability category. The combination of increased salience and hint has changed participants' strategy from classification on the basis of similarity to classification on the basis of likelihood.

It is possible that if the difference in variability is made even more salient, then the hint may in fact have no effect. If the hint acts to draw participants' attention to the difference, then if the difference is already very obvious the hint may have little effect.

Experiment 3

The results in Experiments 1 and 2 suggest that there may be considerable individual differences in relation to the impact of category variability on categorization. However, because each participant produced just one 'critical' response, on each three tasks, there was insufficient data to obtain a clear picture of any such individual variation. We therefore switched to an experimental paradigm

from which a richer picture of an individual's categorization behavior could be obtained. A further motivation for this paradigm was to generalize our previous findings to a category with more than one dimension. Categories defined by two stimulus dimensions are used. The dimensions were the height and width of simple geometric shapes. Two pairs of categories described in the modeling section of this paper were used: pair 1:2 and pair 1:4. The category structure was carefully designed to allow two comparison of the classification of the examples intermediate between the two categories of each pair across the two types of pair, 1:2 and 1:4. Either examples equally distant from the mean of the low variability category (and therefore also equally distant from the mean of the high variability category) could be compared, as in Fried and Holyoak's (1984) experiments, or examples equally distant from the nearest example of the low variability category (and therefore also equally distant from the nearest example of the high variability category) could be compared. This second comparison was possible because the distance between the nearest examples of each category was kept constant across the two conditions. A small number of examples were used so that a MDS solution could be obtained for each set of items, to examine whether the assumption that one category was more variable than the other was true in the participants' psychological spaces. To allow category variability to be easily manipulated without requiring many examples in each category, the category structure was hollow, with the examples distributed in a circle around the category mean. Pilot work showed that very similar generalization gradients are obtained for these circular category structures as are obtained for categories of truly normally distributed examples. Attempting to obtain judgments that can serve as the basis for an MDS analysis using the same participants as in the present study would lead to potential interference between MDS judgments and task

performance. Hence we ran a separate study to conduct this MDS analysis, Experiment 5, reported below.

In this experiment each participant was trained on a pair of categories, and then given transfer examples intermediate between the two categories to allow a generalization gradient to be determined for that structure. The participant will then repeat this process for the other pair of categories. Thus for each participant, there are two generalization gradients, one for each of the two pairs of categories. As described in the modeling section of this paper, the difference in variability between the two categories in each pair is greater for the 1:4 pair than the 1:2 pair. Modeling with the GCM and GRT demonstrated that exemplar and distributional models make different predictions about the proportion of high variability responses to stimuli intermediate between the categories for the 1:2 and 1:4 pairs of categories, when stimuli equally distant from the nearest neighbors of a category are compared across the 1:2 and 1:4 conditions. The exemplar model predicts that the proportion of high variability responses to intermediate stimuli will be lower in the 1:4 condition than the 1:2 condition. The distributional model predicts the opposite – that the proportion of high variability responses to intermediate stimuli will be higher in the 1:4 condition than the 1:2 condition. (Equivalently, the GRT predicts that an example that is equally likely to be classified into either category will be nearer the low variability in the 1:4 condition compared to the 1:2 condition, when the distance is measured relative to the nearest example of each category. The GCM makes the opposite prediction.)

Experiment 3 sets out to find which model describes the behavior of human participants, both at the level of across participant averages, and for individual participants. It is important to consider performance at the level of individual

participants as Maddox (1999) demonstrated that data averaged across participants may not reflect individual participant data, especially when large individual differences exist. Using Monte Carlo simulation, Maddox generated data sets from either GRT or from the GCM. When the GCM was the correct model, averaging has little effect. However, when GRT was the correct model and therefore perfectly describes the generated data, averaging lead to a better fit for the GCM than GRT.

Method

Participants. 32 undergraduates from the University of Warwick participated for course credit, or payment of £5.

Design. The design is fully within participants. Each participant completed two categorization training and transfer tasks. In a training stage participants learned to categorize shapes that varied in height or width into one of two categories on the basis of trial by trial feedback. The shapes of one category were smaller than the shapes of the other category. In the transfer stage participants classified old training examples and new transfer items without feedback. In each categorization task the variability of the two categories differed. In one task the ratio of variability was 1:2, and in the other task the ratio of variability was 1:4. The transfer items in each task were intermediate in height and width between the two categories, and were designed to measure the generalization gradient between the two categories. The order of learning the 1:2 and 1:4 tasks was counterbalanced across participants. To minimize carry over effects, each participant did one task with rectangles of varying height and width, and one task with ellipses of varying height and width. The assignment of shape to variability condition was counterbalanced across participants. The assignment of labels to categories was also counterbalanced. Finally, the assignment of variability to the category of either small or large shapes was also

counterbalanced. That is, for half the participants the small shapes category was more variable than the large shapes category, and for the other half the large shapes category was more variable than the small shapes category. The predictions for classification of the transfer stimuli are outlined in detail in the introduction of this experiment. Briefly, participants classifying on the basis of similarity were expected to classify more transfer items into the low variability category in the 1:4 condition compared to the 1:2 condition. Participants classifying on the basis of likelihood were expected to demonstrate the opposite result.

Stimuli. The heights and widths of stimuli used in the 1:2 condition are given in Figure 7 and stimuli used in the 1:4 condition are illustrated in Figure 8. One unit of height or width corresponds to approximately $1/50^{th}$ of a degree of visual angle. In Figure 7 the low variability category is also the category of shapes with smaller heights and widths. The heights of each example used with the reverse assignment of category mean and variance may be generated by subtracting the height from 500. Similarly for the widths. That is, each category structure is a mirror image of the other in the line $x+y=500$. The same is true of the assignment of variability to location in space in the 1:4 condition.

Apparatus. Stimuli were displayed on a 14" Macintosh Color Display. Responses were collected using labeled keys on a normal qwerty keyboard. The keys Z and X were labeled A and B respectively.

Procedure. The experiment began with some general instructions on the screen of the computer informing participants that they would learn to categorize some shapes into two categories, be tested on those shapes, and then learn to categorize a new set of shapes, and be tested on those. The first stage of the experiment began with specific instructions for the first training phase. Participants

were told that shapes would come up on the screen one after the other, and that they should press one of the labeled keys "A" or "B" depending on which category they though the shape belonged to. They were told that by paying attention to the correct answer displayed on the screen after each response that they could learn the categorization, and that they should try to do this because they would be tested on it later. Each trial started with presentation of a stimulus until the participant responded. Feedback was given on the screen for 1500 ms. The feedback was the correct category label, presented as a letter (A or B) 50 pixels high below the stimulus. The stimulus remained on the screen until the end of the feedback. There was a blank screen pause of 500 ms before the next trial began automatically. The sequence of 100 trials comprised 5 repetition of the 20 training examples. In each repetition the trials were in a random order. At the end of the training session instructions for the transfer session were displayed, telling participants to continue categorizing the shapes as they appeared, despite the absence of feedback. There were 328 transfer trials comprising 8 repetitions of 41 examples. 20 of the 41 examples were the old training examples. The remaining 21 transfer examples were novel examples located in-between the two categories in height width space. Within each repetition the 41 examples were displayed in a random order. The order was different for each repetition. The structure of a trial was the same as in training, except the feedback was omitted. After a participant had responded, the screen was cleared, and the next trial began after a 500 ms pause. When participants had completed the first categorization task they moved on to a second task, which was the same as the first except the category structure was swapped, as was the type of shape.

Results

Average Results. The mean proportion of correct responses in training was 0.93 for both the 1:2 condition and the 1:4 condition. Participants were very accurate in their training classifications. A five way ANOVA (category mean and variance assignment × category label × stage order × rectangle or ellipse × condition) was run to check that none of the counterbalanced factors, or the category structure affected training performance. There was a significant main effect of category mean and variability assignment, $F(1, 16)=6.83$, $p<0.05$, corresponding to a slight increase in the proportion of correct responses for participants assigned to the condition where the category of larger shapes were more variable (0.95 verses 0.91). This effect was not shown in performance on old training items in transfer. There were no other significant main effects (largest $F(1, 16)=4.13$, $p>0.05$). Performance on old training items was also excellent during transfer. The proportion of high variability category responses to old training items is shown in Table 3. A six way ANOVA (category mean and variance assignment × category label × stage order × rectangle or ellipse × condition × category) revealed a main effect of category, $F(1, 16)=6.54$, $p<0.05$. Although performance was high on training examples in test, examples of the low variability category were classified slightly less accurately than examples of the high variability (mean proportion correct 0.89 verses 0.96) category. There were no other significant main effects (largest $F(1, 16)=2.32$, $p>0.05$). This indicates that no counterbalanced factor had a significant effect on old training item classification in transfer. In summary, performance on training items was high, and not greatly influenced by any counterbalanced factor. Further, accuracy was about equal for the low and high variability categories, and also about equal for the 1:2 and 1:4 conditions.

Table 3

Mean proportion of high variability responses across all participants for Experiment 3 split by variability condition category. (Numbers in brackets are standard errors of the means.)

	Variability condition	
Category	1:2	1:4
Low variability	0.11 (0.03)	0.12 (0.03)
High variability	0.95 (0.01)	0.96 (0.01)

It is the performance on the new transfer items that is of interest. Two analyses of these data are reported here. The responses given to each of the 21 new transfer items are collapsed into 7 sets, so that responses to stimuli whose projections onto the line $y=x$ coincide are in the same set. Figure 17 shows a plot of the proportion of high variability responses given to stimuli in each of the 7 sets as a function of their position relative to the means of the two categories. Figure 17 can therefore be thought of as a generalization gradient for the two categories. In both the 1:2 condition and the 1:4 condition, the proportion of high variability responses to test stimuli increased as the test stimuli moved towards the high variability category and away from the low variability category. In the 1:4 condition the proportion of high variability responses was higher than for the 1:2 condition for every set of test stimuli. This description of the results was confirmed by a six way ANOVA (condition × stimulus set × category mean category variance assignment × category label × stage order × rectangle or ellipse). The proportion of high variability responses increased as the test stimulus gets closer to the high variability category, $F(6, 96)=185.77$, $p<0.0005$ (Huynh-Feldt $\varepsilon=0.82$). There were more high variability responses in the 1:4 condition than in the 1:2 condition, $F(1, 16)=10.52$, $p<0.01$. The interaction between stimulus set and condition did not reach significance, $F(6, 96)=1.67$, $p>0.05$ (Huynh-Feldt $\varepsilon=1.00$). There were no other significant main effects (largest $F(1, 16)=1.06$, $p>0.05$), showing that none of the factors counterbalanced across participants affected responding significantly.

By analyzing the results as above, we compared classification of examples that are equally distant from either the mean of the low variability category (or the mean of the high variability category – the two comparisons are equivalent) across the 1:2 and 1:4 condition. However, an example that is equally distant from the low

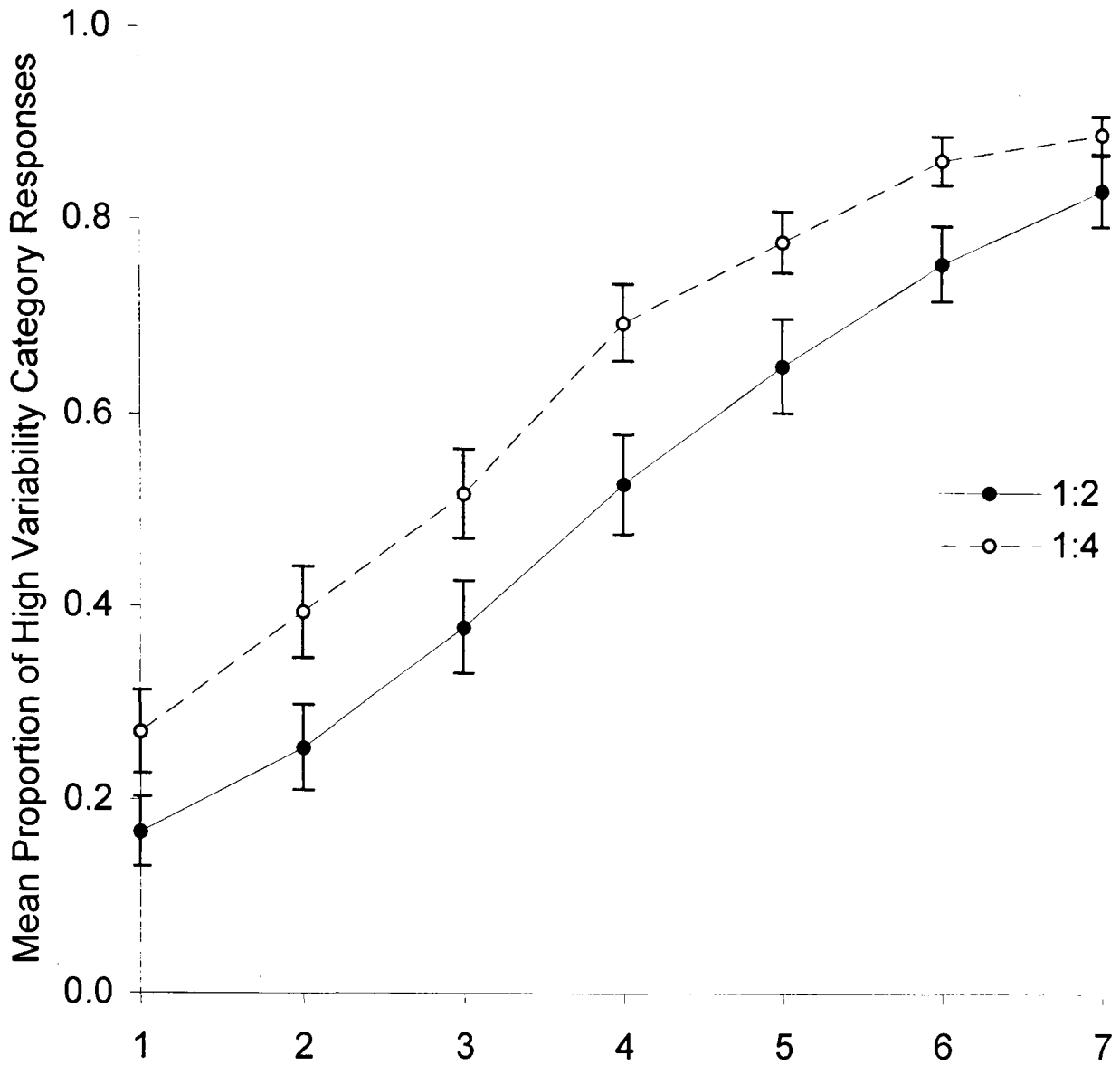


Figure 17. The generalization gradients obtained from Experiment 3 for the pairs of categories 1:2 and 1:4. The absolute position is measured relative to the two category means. (Error bars are standard error of the mean.)

variability category mean in the 1:2 and 1:4 conditions was not equally distant from the nearest example of the low variability category. The following analysis compares examples that are equally distant from the nearest example of the low variability category (or equivalently, examples equally distant from the nearest example of the high variability category) across the two conditions. Numbering the sets of new training stimuli from the one nearest the low variability to the one nearest the high variability category for both conditions, if one compares set n in condition 1:4 with set n+1 in the 1:2 condition, then the comparison is between sets equally distant from the nearest low variability example, or the nearest high variability example. Such comparisons are shown in Figure 18. (If one shifts the 1:2 data in Figure 17 one unit to the left one obtains Figure 18.) Figure 18 can be though of as the generalization gradient between the two categories measured relative to the nearest neighbor of either the low variability category, or the high variability category. (As the distance between the nearest neighbors was the same for both category pairs, it does not matter which nearest example position is measured relative to.) Unsurprisingly we see that, as before, as the test stimulus gets nearer the examples of the high variability category the proportion of high variability responses increases. However now the position of the test set is measured relative to the nearest neighbors of the two categories, rather than relative to the categories' means, there is no difference between the generalization gradients for the two conditions. A further six way ANOVA (condition × stimulus set × category mean category variance assignment × category label × stage order × rectangle or ellipse) confirms this description. There was a main effect of stimulus set, $F(5, 80)=170.01$, $p<0.0005$ (Huynh-Feldt $\varepsilon=0.87$), but now no main effect of condition, $F(1, 16)=0.23$, $p>0.05$, and no stimulus set × condition interaction, $F(5, 80)=0.41$, $p>0.05$ (Huynh-Feldt $\varepsilon=1.00$). There were no

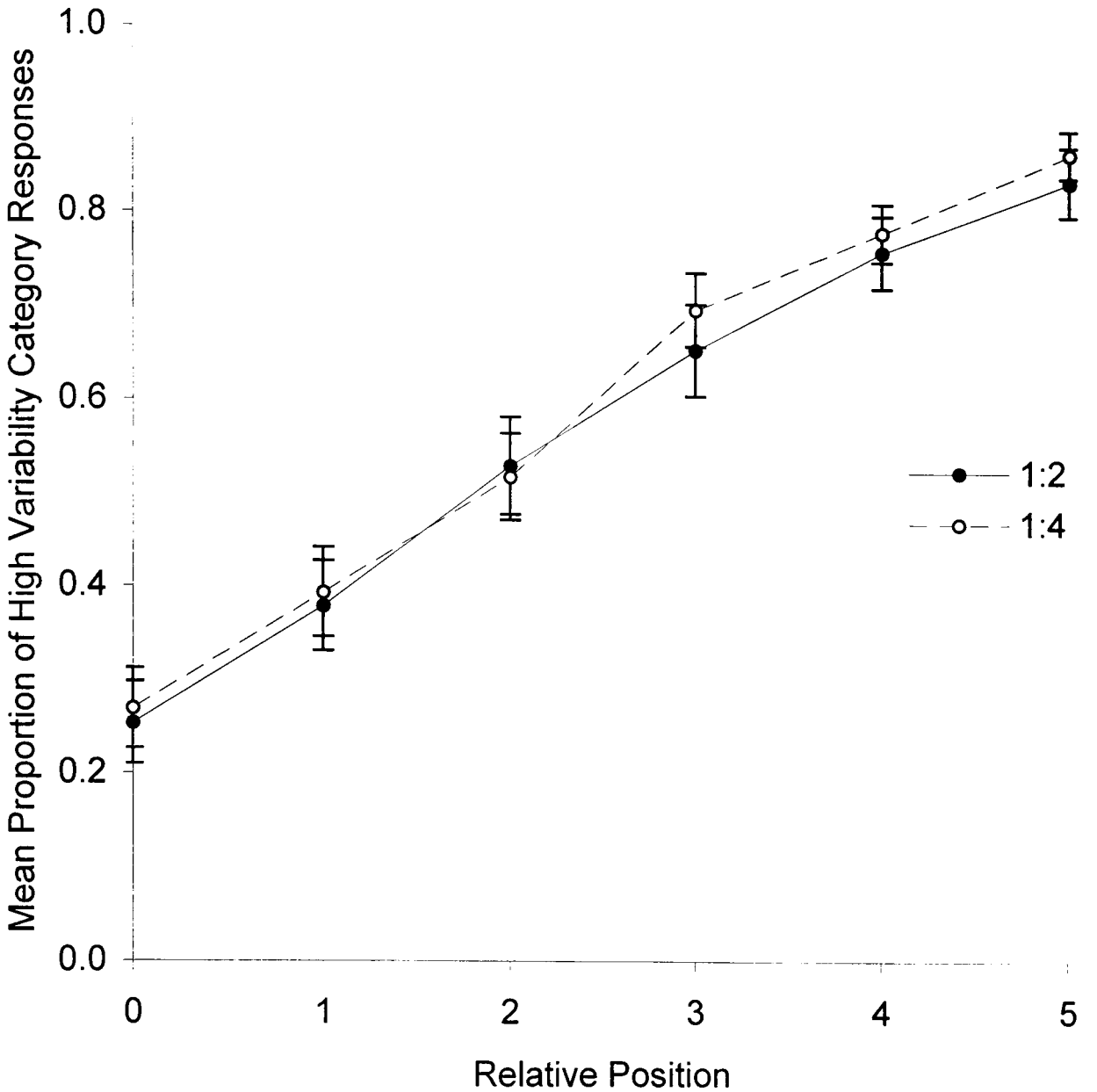


Figure 18. The generalization gradients obtained from Experiment 3 for the pairs of categories 1:2 and 1:4. The relative position is measured relative to the nearest exemplar of the low variability category. (Error bars are standard error of the mean.)

other significant main effects (largest $F(1, 16)=1.19$, $p>0.05$), showing that none of the factors counterbalanced across participants affected responding significantly. It appears that it can be concluded that the difference in generalization gradients between the two conditions is completely explained by the nearest neighbors of each category.

This conclusion would be premature, however, because, as Maddox (1999) has pointed out, averaging across participants does not always produce an accurate account of individual participants' behavior.

Individual Participant Results. When generalization gradients are calculated for individual participants they show many participants show very different gradients for the two conditions. The results averaged across participants do not represent individual performance well. Even when the effect of nearest neighbors is controlled, when the difference in the variability of the two categories is increased from 1:2 to 1:4, 14 participants show an increase in their proportion of high variability responses, and 18 show a decrease. Further, for many of these participants the change is larger than would be expected by chance. A χ^2 analysis was performed for each participant, with the trial as the unit of analysis. A 2×2 (variability condition × response) contingency table was constructed for each participant, and a χ^2 statistic was calculated on the basis of the hypothesis that there should be no difference in the proportion of high variability responses between the two conditions. Yates continuity correction was not used, as there is no reason to expect constant marginal totals, and the expected frequencies were large (Howell, 1997, p. 146). As the assumption that the response on each trial is independent of the response on any other trial is unlikely to be true, the statistic was deflated to account for trials being non-independent (Altham, 1979; see also Tavare & Altham, 1983). 12 of the 32 participants showed a

significant difference between their responding in the two conditions, 7 increasing and 6 decreasing their proportion of high variability responses as the difference in variability between the two conditions increased. The probability of obtaining 13 or more significant differences (i.e., $p<0.05$) by chance is 1.72×10^{-9}, assuming that the number of significant results is binomially distributed ($n=32$, $p=0.05$).

Discussion

Accuracy on the training examples of each category in each condition was high. Averaged across participants, when the difference in variability between two categories is increased, the proportion of high variability responses to intermediate stimuli increases. This result is consistent with the predictions of the GCM and GRT. Of interest here is the result when the presence of nearest neighbors is taken into account. This was done by comparing examples that are equally distant from the nearest neighbor of the low variability category (or equivalently, equally distant from the nearest neighbor of the high variability category) across the two conditions. Averaged across participants, the generalization gradients for the two conditions were virtually identical. This is inconsistent with the predictions of GRT, but is consistent with GCM (when the amount of generalization is small). However, the individual participant data are not well described by the average results. Some participants showed a higher proportion of high variability responses in the condition with greater difference in responding between the two categories. Other participants showed the opposite result. For participants where there was a significant difference between the proportion of high variability responses in the two conditions, half showed higher proportions in the condition where the difference in variability was larger, and half showed higher proportions in the condition where the difference in variability was lower. Neither the GCM nor GRT can explain this result, as each

theory either makes one prediction, or the other. The category structure alone does not determine the strategy used.

There are two possible explanations of the great difference in performance across individual participants. Experiments 1 and 2 demonstrated that participants can change their strategy for classification of intermediate stimuli. It is possible therefore, that participants were each using one of two different strategies. One strategy would be well modeled by the GCM, and the other would be well modeled by normal GRT. The second possibility is that participants' classification of the intermediate stimuli is simply not influenced directly by the training stimuli, as both the GCM and normal GRT predict. Instead, performance on this intermediate region is determined by factors outside of the control of the experiment, perhaps factors investigated in Experiments 1 and 2 – the salience and participants' knowledge of the difference in variability. Participants have no training examples for the region of space between the two categories. It could be, for example, that they make an arbitrary decision that determines performance in this region.

Experiment 4

Both the exemplar and distributional approaches are unable to predict that large variation between individuals demonstrated in Experiment 3. However, if some participants are assumed to apply an exemplar approach, and some a distributional approach this variation may explained. Alternatively, it could be that neither an exemplar approach, nor a distribution approach is a good model of individual participant's data. Experiment 4 aims to discriminate between these two possibilities.

Experiment 4 differs from Experiment 3 only in that the 1:4 pair of categories is replaced with a new pair of categories, 1:2 expanded. All other aspects of the design and procedure are the same. The 1:2 expanded pair of categories differs only

slightly from the 1:2 structure – in the 1:2 expanded pair of categories, the 5 examples of the high variability category that are further from the low variability category are moved to even more extreme points (Figure 14). As demonstrated in the modeling section of this paper, the GCM and normal GRT predict no difference between the generalization gradients for the two pairs of categories. The reason for the lack of difference between the 1:2 condition and the 1:2 expanded pairs of categories, according to the GCM, is that for the model to predict accurate classification of training examples the amount of generalization is so small that distant examples of the high variability category have negligible effect on the classification of the transfer items. Thus moving these examples to more distant locations has no effect. Normal GRT predicts no difference, because the increase in the variability of the high variability category in the 1:2 expanded pair of categories is offset by the category mean moving slightly further away. Therefore, the decision bound for each pair of categories is almost exactly the same, and hence the generalization gradients hardly differ. In summary, both models predict no difference in the classification of the examples intermediate between the two categories for the two pairs of categories. However, the category structures used here are very similar to those used in Experiment 3, so there is good reason to expect replication of the large individual differences.

Method

This experiment only differs in the category structure used. In all other respects it is identical to Experiment 3.

Participants. 32 undergraduates from the University of Warwick participated for course credit, or payment of £5. No participant had taken part in any other experiment in this paper.

Stimuli. The stimuli in the 1:2 condition are the same as in Experiment 3, as shown in Figure 7. The stimuli in Experiment 3's 1:4 condition have been replaced by new stimuli in this experiment. The new stimuli are as for category pair 1:2, except that five stimuli of the high variability category (those most distant from the low variability category) have been moved to more extreme regions of height width space (Figure 14). The transfer stimuli remain unaltered.

Results

Average Results. The mean proportion of correct responses in training was 0.91 for both the 1:2 condition and the 1:2 expanded. Participants were very accurate in their training classifications. A five way ANOVA (category mean category variance assignment × category label × stage order × rectangle or ellipse × condition) was run to check that none of the counterbalanced factors, or the category structure affected training performance. There were no significant main effects and no significant interactions (largest $F(1, 16)=4.18$, $p>0.05$). Performance on old training items was also excellent during transfer. The proportion of high variability category responses to old training items in transfer is shown in Table 4. A six way ANOVA (category mean category variance assignment × category label × stage order × rectangle or ellipse × condition × category) was conducted to examine whether any of the control factors had an effect on performance, and to check that performance on old training items was equal for each category. There was a main effect of learning order, $F(1, 16)=5.84$, $p<0.05$, that corresponds to a 3% accuracy advantage for the group learning the 1:2 condition before the 1:2 expanded condition. Compared to the size of the effect for the new test stimuli, this effect is tiny. Further, an increase in accuracy should sharpen a generalization gradient, but it should not lead to a increase in the proportion of responses to one category, which is what is of

Table 4

Mean proportion of high variability responses across all participants for Experiment 4 split by variability condition category. (Numbers in brackets are standard errors of the means.)

	Variability condition	
Category	1:2	1:2 Expanded
Low variability	0.07 (0.01)	0.10 (0.02)
High variability	0.93 (0.01)	0.93 (0.01)

interest here. There were no other significant main effects (largest $F(1, 16)=1.84$, $p>0.05$). This means no other counterbalanced factor had a significant effect on old training item classification in transfer.

It is the performance on the new transfer items that is of interest. The only difference between the 1:2 and the 1:2 expanded conditions was that some distant examples of the more variable category have been changed to even more extreme locations of stimulus space. Thus each new test example is of equal distance from the nearest example of the low variability category between the two conditions. (That is, the effect of nearest neighbors is controlled across the two conditions without the adjustment required in Experiment 3.) As in the previous experiment's analysis the responses given to each of the 21 new transfer items are collapsed into 7 sets, so that responses to stimuli whose projections onto the line $y=x$ are averaged together. Figure 19 shows a plot of the proportion of high variability responses given to stimuli in each of the 7 sets as a function of their position relative to the means of the two categories. For both the 1:2 pair and the 1:2 expanded pair, the proportion of high variability responses to test stimuli increases as the test stimuli moves towards the high variability category and away from the low variability category. There is almost no difference between the proportion of high variability responses in the 1:2 condition and the 1:2 expanded condition. This description of the results is confirmed by a six way ANOVA (condition × stimulus set × category mean category variance assignment × category label × stage order × rectangle or ellipse). The proportion of high variability responses increases as the test stimulus gets closer to the high variability category, $F(6, 96)=277.20$, $p<0.0005$ (Huynh-Feldt $\varepsilon=1.00$). There is no significant difference between the proportion of high variability responses in the 1:2 expanded condition and in the 1:2 condition, $F(1, 16)=0.25$,

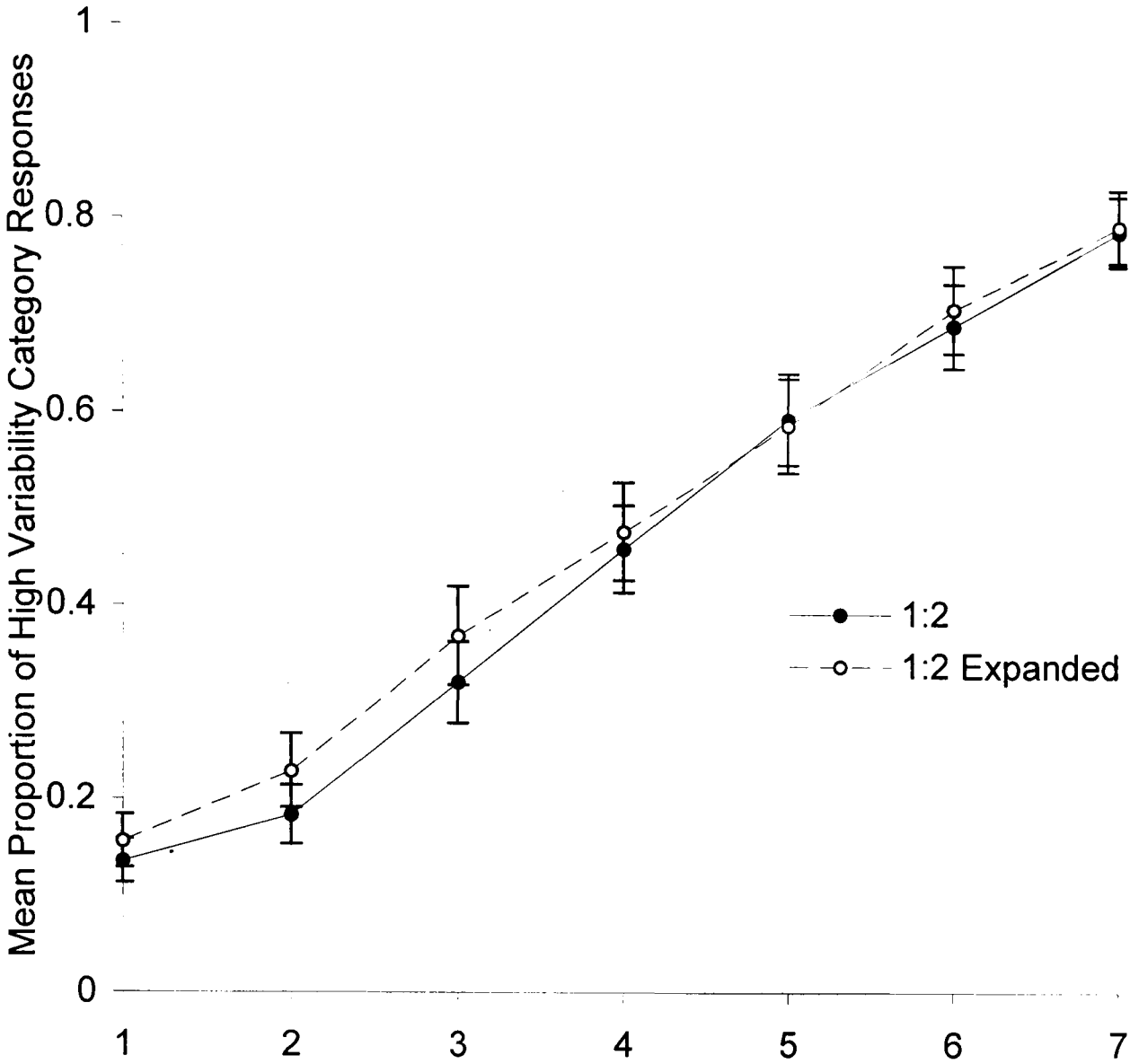


Figure 19. The generalization gradients obtained from Experiment 4 for the pairs of categories 1:2 and 1:2 expanded. The absolute position is measured relative to the two category means, and is equivalent to the position measured relative to the nearest exemplar of each category. (Error bars are standard error of the mean.)

$p>0.05$. The interaction between stimulus set and condition did not reach significance, $F(6, 96)=0.61$, $p>0.05$ (Huynh-Feldt $\varepsilon=0.74$). None of the counterbalanced factors had a significant effect (largest $F(1, 16)=3.88$, $p>0.05$), showing that none of the factors counterbalanced across participants affected responding significantly.

Individual Participant Results. The results averaged across participants do not represent individual performance well. As for Experiment 3, when generalization gradients are calculated for individual participants they show many participants have very different gradients for the two conditions. When the distant examples of the more variable category are moved to be more extreme points 8 participants show an increase in their proportion of high variability responses, whereas the remaining 24 show a decrease. Further, for many of these participants the change was larger than would be expected by chance. As before a χ^2 analysis was performed for each participant, with the trial as the unit of analysis. 19 participants showed a significant difference between their responding in the two conditions, 4 increasing and 15 decreasing their proportion of high variability responses as the difference in variability between the two conditions increased. The probability of obtaining 19 or more significant differences (i.e., $p<0.05$) by chance, if one assumes there is no difference between the proportion of high variability response in the two conditions, is $p=3.52\times10^{-17}$, assuming that the number of significant results is binomially distributed ($n=32$, $p=0.05$).

Discussion

Moving the distant examples of the high variability category to even more distant locations did not alter the generalization gradient obtained from averaged participants' data. This result is consistent with the predictions of the GCM and

normal GRT. As in the previous experiment, individual participant data is not well described by the average data. Out of the 32 participants tested, 19 showed a significant difference in the proportion of high variability responses in the two conditions. The remaining participants showed differences that failed to reach significance. For more than half of the participants, moving the distant examples had an effect on their performance on the transfer items. Both the GCM and normal GRT are unable to account for this result.

These results suggest that the second account of the results of Experiment 3 should be preferred – that the location of the generalization gradient (either near the low variability category or near the high variability category) is not determined by the category structure in the way that exemplar and distributional account predict it should be. The first possibility – that some participants use a GCM strategy, and others a GRT strategy – can be rejected, as at the level of individual participants neither strategy can account for the majority of participants' results in this experiment.

Experiment 5

It is not always the case that a category structure in psychological space reflects the structure of the category in the experimenter's choice of physical space (e.g., Palmeri & Nosofsky, in press). If the category structure of stimulus sets used in Experiments 3 and 4 in psychological space of the did not reflect the category structure in physical space, then the conclusions drawn from Experiments 3 and 4 may be incorrect. To investigate this possibility multidimensional scaling solutions were obtained for the three category structures used in Experiments 3 and 4 (Nosofsky, 1986).

Method

Participants. 12 University of Warwick Undergraduates participated for payment of £5. None of the participants had participated in any of the previous experiments in this chapter.

Design. The 12 participants were randomly assigned to one of three stimulus sets, with the constraint that there were 4 participants assigned to each set. The stimuli are described in detail in Experiment 3 (1:2 and 1:4 category pairs) and Experiment 4 (pair 1:2 expanded). Within each group of 4 participants assignment of category mean to the more and less variable categories was counterbalanced, as was the instantiation of the stimuli as either ellipses or rectangles. 22 stimuli from each set were used. The first 20 were the 10 training examples from each of the two categories. The MDS solutions obtained for these examples would allow the assumption that the variability of the categories in physical space was also true in psychological space. The last two were the two new central transfer examples. These examples were included to check that the transfer examples occupied a point intermediate between the two categories in psychological space. Ever possible pair of the 22 stimuli (484 pairs) was presented to participants.

Procedure. Participants were seated in front of the computer, and the position of the screen and keyboard adjusted as necessary. The experiment began with instructions informing participants that they would see a number of pairs of stimuli on the screen, and that for each pair they should make a judgement as to how similar the pair were, and response using the scale provided. The experimenter checked the participant understood the instructions, and the experiment began. On each trial a pair of stimuli from the stimulus set the participant was assigned to appeared next to one another on the screen. One stimulus was on the left side of the screen, and the

other on the right side of the screen. The shapes remained on the screen until the participant pressed one of the number keys 1-9 along the top of the keyboard. The screen was then cleared, and there was a 500 ms blank before the next trial began. Each pair was displayed once, and the pairs were displayed in a random order for each participant.

Results

The INDSCAL multidimensional scaling model (Carroll & Wish, 1974a; Shepard, 1980) was used to derive solutions for the three stimulus sets. The INDSCAL model takes as its input an average similarity matrix for each participant and produces an MDS solution common to all participants, together with a weight set for each participant. In other words, the procedure produces a configuration of stimuli common to all participants that best describes each participants similarity matrix, under the assumptions that the relative contribution of each dimension of the solution may vary across participants, and that each participant may use the response scale differently. For each stimulus set a 6 dimensional solution was produced. A large number of dimensions was chosen to preserve the information in the confusion data that would be lost with a smaller number of dimensions. The overall fit of the INDSCAL model was quite good for each of the stimulus sets, as reflected by the low Stress values and the high r^2 values in Table 5. The solutions themselves are given in the Appendix for the stimulus sets 1:2, 1:4 and 1:2 expanded.

In Experiments 3 and 4, a crucial assumption is that the difference in variability between the two categories is bigger in the 1:4 and 1:2 expanded category pairs than the 1:2 category pair. To test this assumption the within category distance was calculated from the MDS solution for each category (Table 6). First, the co-ordinates of each stimulus in the MDS space were used to calculate a matrix of inter-

Table 5

Stress and $\underline{r}^2$ values for the INDSCAL MDS solutions obtained for the three different category structures.

	1:2	1:4	1:2 Expanded
Stress	0.105	0.133	0.105
$\underline{r}^2$	0.840	0.674	0.836

Table 6

The mean inter-stimulus distance within a category derived from the MDS solution. (Comparisons across rows are meaningless, as the value of one unit is arbitrary and different in each solution.)

	1:2	1:4	1:2 Expanded
Low variability category	2.42	2.44	2.36
High variability category	3.03	3.29	3.32

stimulus distances. Second, the inter-stimulus distances between items of the same category were averaged together for each category. In line with the assumption, the ratio of the mean within category distances is larger for the 1:4 pair than the 1:2 pair, and approximately equal to the ratio for the 1:2 expanded pair. (The reader should note that as the units of the MDS space are arbitrary for each solution, comparisons across the rows of Table 6 are meaningless. One can only compared distances within a single solution.)

A further assumption in Experiment 4 was that, according to the GCM, the more distant examples of the high variability category did not contribute significantly to the summed similarity of the transfer examples to the examples of the high variability. That is, only the five examples of the high variability category that are nearest the transfer items significantly affect the GCM's classification of the transfer examples. Table 7 shows these summed similarities of the transfer examples for a range of values of the c parameter. The summed similarities were calculated using the MDS co-ordinates for both the 1:2 and 1:2 expanded pairs. When the c parameter is large enough that the accuracy on the training items is equivalent to that achieved by participants in Experiments 3 and 4, the contribution of the distant examples of the high variability category to the total summed similarity of the transfer examples to the high variability category is very small compared to the contribution of the near examples of the high variability category. This is true for both the 1:2 and 1:2 expanded pairs. The results in Table 7 are generated from the GCM with a Euclidean distance metric and a Gaussian similarity function, but the same pattern is obtained if a city block distance metric and an exponential similarity function are used.

Table 7

The GCM's predictions for summed similarity for different c parameters. (These results are for a Euclidean Gaussian model, but a similar pattern is obtained for a city block exponential model.)

		Mean accuracy		Summed similarity to			
	c	low variability	high variability	low variability category	high variability category	high variability near exemplars	high variability far exemplars
1:2							
	0.3	0.69	0.64	3.95	4.01	2.62	1.39
	0.4	0.82	0.76	2.04	2.30	1.78	0.52
	0.5	0.92	0.88	0.96	1.35	1.20	0.15
	0.6	0.97	0.96	0.43	0.85	0.82	0.03
	0.7	0.99	0.99	0.19	0.57	0.57	0.01
1:2 expanded							
	0.4	0.70	0.61	2.54	1.87	1.40	0.47
	0.5	0.77	0.67	1.44	1.02	0.88	0.13
	0.6	0.82	0.73	0.84	0.64	0.61	0.03
	0.7	0.84	0.77	0.51	0.48	0.47	0.01
	0.8	0.85	0.80	0.32	0.39	0.39	0.00

Discussion

In Experiments 3 and 4 it was assumed that participants' representation of the category structure would reflect the physical category structure. Experiment 5 used a MDS procedure to recover participants' representation from pair wise similarity judgments. Two assumptions were tested. The first assumption was crucial to Experiment 3 and Experiment 4 – that the high variability category was more variable than the low variability category. The average between example distance was higher for conditions 1:4 and 1:2 expanded, than it was for condition 1:2, confirming this first assumption. The second assumption was crucial to Experiment 4 – that the distant examples of the high variability category in the 1:2 and 1:2 expanded conditions made negligible contribution to the summed similarity of each transfer example to that category. Modeling with the GCM confirmed this to be true for the range of c parameters that would produce accuracy on training examples approximating that obtained by participants in Experiments 3 and 4. In summary, the MDS solutions obtained for each stimulus set confirm that the assumptions made about the representation of the sets in Experiment 3 and 4 were true.

The GCM and GRT were not fitted to the individual participant data from Experiments 3 and 4 because fitting the models would provide no useful information. The models were used to generate qualitative predictions about how altering the relative variability of two categories should alter performance. Simply seeing which model fits the data best therefore seems inappropriate.

General Discussion

The experiments presented in this paper investigate whether performance in simple perceptual categorizations is based on similarity to stored category examples or likelihood in relation to a probability distribution inferred from the data. Modeling

using an exemplar model (the GCM, Nosofsky, 1986) and a distributional model (GRT, Ashby & Townsend, 1986) demonstrated that the two accounts differ in their predictions for the classification of an example exactly intermediate between the nearest examples of two categories that differ in variability. The exemplar model predicted classification of the intermediate example into the more similar, lower variability category, but the distributional model predicted classification into the more likely, higher variability category. Further, it was demonstrated that the exemplar and distributional models make opposite predictions about the effect of increasing the relative variability of the two categories on classification of intermediate examples. The exemplar model predicted that the probability of classifying an intermediate example into the high variability category would decrease as the difference in variability increased. At odds with this prediction, the distributional model predicted that the probability of classifying an intermediate example into the high variability category would be increased as the difference in variability increased.

Experiment 1 showed that on average participants classified a stimulus half way between the nearest neighbor from each category into the lower variability category, whose examples are more similar to the critical stimulus. This is only consistent with similarity based classification. Experiments 2 demonstrate that increasing the salience of the difference in variability between two categories, by presenting examples simultaneously rather than sequentially, encourages participants use of a likelihood strategy rather than a similarity strategy. Further, when the variability is more salient, drawing participants' attention to the difference in variability between the two categories further encourages use of a likelihood strategy. Experiment 3 demonstrated that individual participants varied greatly in the

change in the categorization of stimuli intermediate between the two categories as the relative variability of the pair of categories was increased. Some participants showed an increase in high variability responses, consistent with the predictions of normal GRT, and others showed a decrease, consistent with the predictions of the GCM. This result suggests that it is possible that people's behavior may be strategic – in the sense that it is determined by their understanding of the nature and demands of the task, rather than tapping into some basic aspect of cognition. The best construal for GCM and normal GRT would be that both kinds of mechanism are available to people, and they can choose between them. However this seems to involve the cognitive system in unnecessary duplication, given that the two approaches produce extremely similar answers under almost all circumstances. This possibility is eliminated by the results of Experiment 4. Experiment 4 replicated the results of Experiment 3 using two pairs of categories where both exemplar and distributional models were constrained to predict no change in the proportion of high variability responses to intermediate stimuli as the relative variability of the two categories was increased. The majority of participants showed a significant change in the proportion of high variability responses to intermediate stimuli as the relative variability of the two categories was increased, at odds with the predictions of both the GCM and normal GRT. At the level of data averaged across participants these differences between the two differing variability conditions disappear. That the true form of individual participant data is obscured by averaging further illustrates the dangers of averaging across participants (Maddox, 1999).

Exemplar and distributional models can be though of as lying at opposite ends of a continuum of finite mixture models, where the number of distributions used to represent a category varies from one, as in GRT, to the number of examples

of that category, as in the GCM (Ashby & Alfonso-Reese, 1995; Rosseel, 1996). Also contained in this continuum are back propagation networks with sigmoidal activation functions (Rumelhart et al., 1986) and radial basis functions (Moody & Darken, 1989). With small numbers of hidden units (and hence small numbers of free parameters in relation to the size of the data to be modeled), neural networks are analogous to distributional models, because they can only learn data with a particular distributional structure. But if the number of hidden units is large in relation to the amount of data to be learned, then the neural network becomes analogous to an exemplar model, in that any data set can be modeled, whatever its structure, simply learning each piece of data (each example) by rote. The results of Experiment 4 present a serious challenge to unitary accounts of this kind that assume that categorization is achieved by a mechanism at some point along the continuum between distributional and exemplar models, and that this point on the continuum is invariant within and between individuals. Thus, these findings challenge standard formulations of both exemplar and distributional accounts of categorization.

In this paper the issue of category biases has been neglected. There is no doubt that the findings in this paper can all be explained by either the GCM or GRT, if the biases for each category are used as a free parameter in the models. However, allowing the bias to vary as a free parameter provides no predictive power as to how or why there is sensitivity to variability.

As Ashby and Waldron (1999) note, a number of models in the categorization literature assume a low resolution map of perceptual space where regions are mapped onto category labels, for example; the grid model (Ashby & Maddox, 1989); the covering version of ALCOVE (Kruschke, 1992); Anderson's (1991) rational model; the striatal pattern classifier (Ashby & Waldron, 1999). These

models of categorization may leave regions of perceptual space unlabelled. Thus, unlike exemplar and distributional models, these models make no prediction about the classification of stimuli in this region of space. It would seem that this is more appropriate as Experiments 3 and 4 demonstrate that performance in regions of space where participants have no training examples is not specified directly by the other training examples. Both the GCM and normal GRT predict that performance should be specified by those examples. Further, Experiment 4 demonstrates performance in this region that cannot be accounted for by either model. Instead, participants may have to make a conscious decision about the classification of these intermediate examples, that may be influenced by a number of factors. Two such factors were investigated in Experiments 1 and 2 – knowledge about the difference in variability, and the salience of that variability difference during category learning. The point is that these mapping models of categorization are not challenged by the results in this paper because they make no claims about the categorization of new stimuli from regions of perceptual space where participants have no prior examples.

Similarly decision bound models of categorization may be adapted to offer a potential account of these results. Decision bound models are closely related to normal GRT, except participants are assumed to estimate the parameters of the decision bound directly, rather than calculating the bound from the inferred normal distributions used to represent each category. In these experiments, there is a large, empty region between the two categories, where participants have no training data. Therefore, there is a large set of possible decision bounds that participants could use, if they are estimating the bound directly. Decision bound theory could account for the results of Experiments 3 and 4, where participants are shown to vary greatly in their classification of stimuli in this region between the categories, if it is assumed

that some participants have a bound near the low variability category, and some have a bound near the high variability category. However, decision bound theory does not provide a mechanism for explaining how the location of this bound might be influenced by knowledge and salience of the differences in variability, as demonstrated in Experiments 1 and 2.

In conclusion, Experiments 1 and 2 demonstrate that participants' strategy for classifying an ambiguous stimulus can be shifted from classification into the most similar category to classification into the most likely category by increasing the salience of the difference in variability of two categories, and instructing participants of this difference. At the level of individual participants, Experiment 3 shows that participants responding to examples intermediate between the two categories is altered when the relative variability of the two categories is manipulated. Experiment 4 replicates this finding and demonstrates that the sensitivity of the majority of participants cannot be explained either by the GCM or normal GRT. This presents a serious challenge to both exemplar and distributional models of classification as unitary models of categorization behavior.

Appendix

INDSCAL MDS Solutions for the Stimulus Sets Used in Experiments 3 and 4

Table A1

Six-dimensional INDSCAL MDS solution for the 1:2 stimulus set.

	Dimension					
Stimulus	1	2	3	4	5	6
low var 1	1.1313	-0.0519	1.0703	-0.4901	1.4448	-0.1055
low var 2	1.2324	-0.325	0.5961	-1.1343	-0.0373	0.7989
low var 3	0.8673	0.3734	0.8467	-1.1412	-0.6675	1.192
low var 4	0.323	1.007	0.6303	-1.5127	-1.1217	0.5717
low var 5	0.3124	1.54	0.9577	-0.909	-0.2145	1.0218
low var 6	0.17	1.6015	0.8589	-0.721	1.1729	-0.1855
low var 7	1.0339	1.0387	-0.4511	1.5906	1.0489	-0.4567
low var 8	0.7209	1.3503	-0.2525	1.4459	1.2334	-0.7492
low var 9	0.6251	1.5449	-0.658	0.9391	0.7165	-1.1978
low var 10	0.7913	1.1688	0.0176	0.1922	0.5765	-1.8355
high var 1	-1.384	-0.7604	0.8537	-0.3103	-0.4839	-2.0779
high var 2	-0.7239	-1.1937	1.9486	1.4877	-0.8957	0.0073
high var 3	-1.273	-0.7435	1.1998	1.7908	-1.1482	0.5684
high var 4	-1.4027	-0.8062	0.785	1.5769	-0.2478	1.0618
high var 5	-1.3788	-0.6126	-0.4704	-0.092	1.9039	1.5562
high var 6	-0.963	-1.1319	-0.7192	-0.8739	1.7024	-0.295

(table continues)

	dimension					
stimulus	1	2	3	4	5	6
high var 7	0.527	-1.5224	-1.1487	-0.8232	-0.4401	-0.3111
high var 8	-0.3393	-0.3309	-1.9335	-0.6681	-1.1066	0.4218
high var 9	-1.1125	-0.0414	-1.5244	-0.0142	-0.9323	0.9834
high var 10	-1.5223	-0.2948	-0.7577	-0.0445	-0.8724	-1.7244
transfer 1	1.2603	-0.6492	-0.7758	0.0202	-0.8028	0.6715
transfer 2	1.1048	-1.1606	-1.0731	-0.3088	-0.8285	0.0839

Table A2

Six-dimensional INDSCAL MDS solution for the 1:4 stimulus set.

	Dimension					
Stimulus	1	2	3	4	5	6
low var 1	-0.9074	-0.2336	-1.0138	-1.3008	0.7656	0.3594
low var 2	-0.9023	1.1771	1.1679	-0.4735	0.3761	0.01
low var 3	-0.714	-0.0803	1.4536	-0.5601	0.2836	1.6829
low var 4	-0.8272	-0.7567	1.6985	-0.312	0.7208	0.501
low var 5	-1.0707	-1.126	1.1007	0.3895	-0.1905	0.0941
low var 6	-1.0211	-0.4755	0.1353	1.2848	-1.3167	-0.581
low var 7	-1.1404	-0.3238	-1.0869	0.7412	-1.3613	-0.1367
low var 8	-0.9403	-1.2698	-1.3359	0.0869	0.429	0.2425
low var 9	-0.9764	-1.055	-1.1969	-0.5871	0.5838	-0.7721
low var 10	-0.8659	-1.0688	-1.0136	0.1762	0.2822	-1.4406
high var 1	1.0748	0.3405	-0.8869	0.7739	1.4883	2.0804
high var 2	0.602	1.2915	-0.5218	1.6711	1.9011	0.488
high var 3	1.1597	0.2779	-0.098	2.47	0.3868	0.0673
high var 4	1.1755	0.2617	0.9886	0.929	0.9103	-2.0204
high var 5	1.4831	-1.1161	1.6175	0.0695	-0.5963	-0.1139
high var 6	0.9885	-0.5835	0.5763	0.2473	-1.8632	1.4598
high var 7	0.0053	1.862	-0.3559	-0.013	-1.5228	-0.8713
high var 8	0.8593	-0.1194	-1.4714	-1.1879	-1.4901	0.8369

(table continues)

	Dimension					
Stimulus	1	2	3	4	5	6
high var 9	1.3894	-0.4844	0.4899	-1.4865	-0.0108	-1.3433
high var 10	1.5772	-0.3089	-0.589	-0.8029	0.9651	-1.1916
transfer 1	-0.653	1.8694	0.0638	-1.1381	-0.5507	0.3002
transfer 2	-0.2961	1.9216	0.278	-0.9775	-0.1902	0.3486

Table A3

Six-dimensional INDSCAL MDS solution for the 1:2 expanded stimulus set.

	Dimension					
Stimulus	1	2	3	4	5	6
low var 1	1.0074	0.0907	-0.3944	-1.4681	0.3764	0.3003
low var 2	1.0761	-0.1072	-1.2638	-0.1858	0.3137	0.3984
low var 3	0.886	0.6738	-0.1884	1.4572	0.7736	0.6412
low var 4	0.5997	0.6961	0.8727	1.324	-1.3753	-0.6273
low var 5	0.6477	0.8309	1.2937	1.6252	-0.5674	-0.0665
low var 6	0.8634	0.8297	1.1831	1.0043	0.4278	0.8136
low var 7	0.8881	0.0688	1.5574	-0.1599	0.122	1.1974
low var 8	0.7363	-0.5876	1.4625	-1.178	-0.4369	0.5288
low var 9	0.8136	-0.2835	1.3898	-1.3947	-0.4881	-0.326
low var 10	0.8324	-0.4044	0.7372	-1.9298	-0.2202	0.0372
high var 1	-1.2549	0.0201	0.4873	-0.3183	1.4426	-2.3872
high var 2	-0.4826	1.0958	-0.5699	-0.7499	2.8288	-0.5581
high var 3	-1.3992	2.2358	-0.2281	-0.784	0.0347	-0.5271
high var 4	-1.661	0.9282	-0.4387	-0.2043	-0.419	1.8261
high var 5	-1.486	0.2069	-0.5905	-0.477	-0.8744	1.9551
high var 6	-0.8982	0.5871	-0.8073	0.0948	-2.1842	-1.1851
high var 7	0.5299	-0.6377	-1.7302	0.5358	-0.9913	-0.6917
high var 8	-0.5235	-2.0711	-0.2904	-0.2917	0.2603	-1.5317

(table continues)

	Dimension					
Stimulus	1	2	3	4	5	6
high var 9	-1.2497	-2.0093	-0.1608	0.2023	0.0369	0.5543
high var 10	-1.4281	-1.6479	0.5011	0.9449	1.0164	0.0069
transfer 1	0.8117	-0.396	-1.1002	1.4412	0.4031	0.2753
transfer 2	0.6909	-0.1192	-1.722	0.5117	-0.4793	-0.634

Chapter 3

Identification and Categorization of Simple Perceptual Stimuli: A Memory and Contrast Model

Abstract

Categorization research typically assumes that the cognitive system has access to an accurate representation of the absolute magnitudes of the properties of stimuli. and that this information is used in reaching a categorization decision. However, research on identification of simple perceptual stimuli suggests people have very poor representations of absolute magnitude information, and shows that judgments about absolute magnitude are strongly influenced by preceding material. The experiments presented here investigate such sequence effects in categorization tasks. Strong sequence effects were found. Classification of a borderline stimulus was more accurate when preceded by a distant member of the opposite category than by a distant member of the same category. It is shown that category contrast cannot be accounted for by modified exemplar or modified decision bound models of categorization. An alternative memory and contrast model is presented, and is shown to account for the results.

Identification and Categorization of Simple Perceptual Stimuli: A Memory and Contrast Model

Categorization models are often divided into two general classes, each including a wide range of specific accounts: parametric Thurstonian decision-bound models (e.g., Ashby, 1992; Ashby & Perrin, 1988), and non-parametric exemplar models (e.g., Medin & Schaffer, 1978; Nosofsky, 1986). However, all extant models of categorization assume that items can be represented in terms of their (more or less noisy) absolute location in a multidimensional space. This absolute location information is then assumed to be used in the decision process (either directly, as in exemplar models, or indirectly, in relation to decision bounds). Thus two key assumptions are (a) that the absolute location of a stimulus in multidimensional space is available when a categorization or identification decision is made, and (b) that absolute location information provides the sole basis for categorization decisions. Formal models of categorization and identification based on these assumptions have a long history of successful application to a wide range of experimental paradigms (e.g., see Estes, 1994 for a review). Here we question both of the assumptions.

Difficulty in Determining Absolute Magnitudes

Regarding the first assumption above, participants often have difficulty making accurate estimates of the absolute values of stimuli along simple perceptual dimensions, particularly in the absence of contextual information. For example, in a series of classic experiments by Garner (1954) participants were completely unable to determine which of three tones was half as loud as a reference loudness. Instead, participants' judgments were entirely influenced by the range of the three tones (see also Helson, 1964). Absolute magnitude judgments are very coarse – people are only

able to divide items reliably into about five 'bins' on a single dimension, however broad those bins are made (cf. Miller, 1956). Laming (1997) provided extensive discussion of these and other similar findings. Of course, more or less accurate determination of absolute magnitude is often possible; the information may be deduced from perception of the relative magnitude of the stimulus in comparison to an amalgam of reference or context stimuli. If this is the case, then absence of appropriate reference stimuli will pose difficulty. To consider an extreme example, suppose one is listening to tones over headphones and is required to make a simple binary loud/quiet categorization. To the extent that one is unable, in the absence of feedback or a reference loudness, accurately to judge absolute loudness the task will be impossible. Some absolute magnitude information will be available when different stimuli are used and contextual magnitudes are available, but this simply points to the importance of context and feedback as the basis for decision making.

Note that the success of current models of categorization can not be taken as reason to ignore this problem: The use of random or controlled trial orders in almost all categorization experiments, followed by averaging over all stimuli of the same type, discards the very information about sequential context that may provide the true basis for categorization. Thus a primary aim of the research presented here is to examine sequence effects in categorization.

What Information is Used in Categorization?

The second key assumption embodied in many current categorization models is that categorization decisions are based on the (perceived or inferred) location of items in multidimensional space. Note that this issue of information use can be examined separately from the related issue of information availability discussed above; even if accurate information about absolute magnitude is available, whether

directly or indirectly, that information need not be used in identification and categorization decisions.

Much research demonstrates that the absolute identification of stimuli is heavily context dependent in that the response on trial n is influenced by the stimulus and response on trial n-1. In an absolute identification paradigm, participants are presented with stimuli that vary along a (normally uni-dimensional) psychological continuum (e.g., sounds that vary in amplitude, or lines of different lengths). Each stimulus is associated with a unique response. Normally the responses are arranged such that their order corresponds to the order of the stimuli in the psychological space. For example, if 10 lines lengths are used, of 1 cm, 2 cm, … , and 10 cm, and 10 numbers for the responses, 1, 2, … , and 10, each line length would be associated with a single number. The 1 cm line could be associated with response 1, the 2cm line with response 2, and so on. On presentation of a stimulus, a participant is required to identify the unique response for that stimulus. One crucial finding is that the response given to the current stimulus is assimilated to the immediately preceding stimulus (Lacouture, 1997; Ward & Lockhead, 1970; Ward & Lockhead, 1971). In other words, participants are systematically biased to respond as if the current stimulus is nearer the previous stimulus than it actually is. For example, if participants get item 1 followed by item 6, they will show a tendency to respond 5 instead of 6. The effect of stimuli further back in the sequence is the opposite – a contrast effect (Lacouture, 1997; Ward & Lockhead, 1970; Ward & Lockhead, 1971). Thus identification decisions depend on recent previous trials. Of course categorization decisions are not thought to be independent of previous trials, as it is precisely these trials that provide the information the categorization is based on. But exemplar models do typically assume this information is not biased by the local

sequential context provided by recent trials (for an account in terms of criterion-shifting within a Thurstonian framework, see Treisman, 1985; Treisman & Williams, 1984).

Mori (1989) demonstrated that in absolute identification of uni-dimensional stimuli (e.g., frequencies or amplitudes), the information used by the decision process was limited to about 2.5 bits, and that this information was predicted almost completely by the current stimulus, the previous stimulus, and the previous response. This suggests the intriguing possibility that the relation between recent successive trials and the current trial may solely determine the decision making process. For example, in a binary categorization task an extreme possibility would be that each decision is made entirely on the basis of the perceived difference between the current and the previous stimulus, i.e., with no reference to the absolute magnitude of the stimulus. Note that this is a stronger claim than simply that decisions on successive trials are not independent. The claim is that the difference between stimulus on trial n-1 and the stimulus on trial n determines the response given to stimulus on trial n. Participants would respond with the same category label as on the previous trial if there is a small difference between the two stimuli, and a different label if the difference is large. Indeed such a strategy would be the only one available to participants in the absence of absolute magnitude information. In more realistic situations, where partial absolute magnitude information is likely to be available, such a strategy is not likely to be used exclusively. However for purposes of explication we consider the extreme possibility that subjects use only this memory and contrast (MAC) strategy. (A related concept is the "Bypass Rule": Krueger & Shapiro, 1981.)

The Memory and Contrast Strategy

Although intuition suggests that a MAC strategy will lead to very poor performance, preliminary modeling work indicates that strategies of this type can be surprisingly successful. For example, consider a binary categorization for uni-dimensional stimuli, where it is assumed that participants only have access to the magnitude and the direction of the difference between the current trial and the previous trial. By optimizing the size of the difference needed to give a switch in categorization response participants can achieve an accuracy of 85% in categorizing examples in a randomly ordered sequence of trials (independent of the number of items). Such a model works by taking advantage of the correlation that exists between magnitude differences and category shifts when uni-dimensionally varying stimuli are involved. Consider the case depicted in Figure 20 of ten stimuli, equidistant from one another along a single dimension (such as loudness or pitch), divided into two categories. If a correct category A response is given to stimulus 1 on trial n-1, and there is a large positive dimensional shift up the scale to the stimulus on trial n, the large positive shift will be accompanied by a shift to a category B response. A small shift, in contrast, is more likely to represent a within-category shift. An adaptive system could select the optimal shift size over which a change in responding should ensue. Although surprisingly successful, at least in the uni-dimensional case, this strategy will clearly lead to characteristic errors under particular circumstances. For example, if item 1 is followed by item 5 the large inter-trial difference will lead to an erroneous shift in response from category A to category B. In other words, large within category shifts will induce errors. Compare this to a large between category shift, for example item 10 preceding item 5. The large shift will again cause a switch in response, this time correctly.

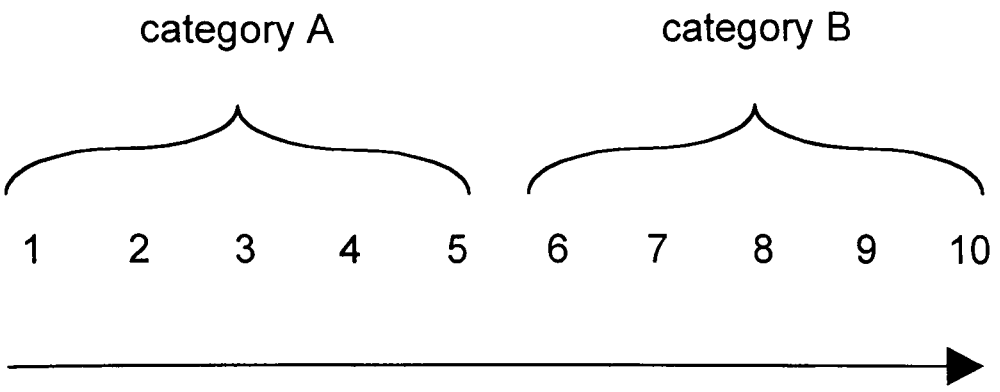


Figure 20. Ten stimuli distributed evenly along a single psychological dimension divided into two categories.

Traditional exemplar models make opposite predictions. Exemplar models can be adapted to predict sequence effects by assuming that more recent exemplars are more available in memory and/or weighted more heavily in the subsequent decision process (e.g., via the M parameter of Nosofsky & Palmeri, 1997; see also Elliott & Anderson, 1995). In exemplar models the probability of responding with a given category label is given by the ratio of the summed similarity to that category, divided by the summed similarity to all contending categories (i.e., in terms of Luce's (1959) choice model). Therefore the probability of responding with a given category can only be increased if exemplars of the same category are weighted more heavily in decision making, as when they have occurred very recently. The consequence of this is that when the item on the preceding trial is from the same category this must always lead to a greater tendency to respond with that category label, relative to the case where the previous stimulus was from the other category. This is the opposite prediction to that made by the MAC model described above.

Modeling

To support the intuitive argument above, categorization performance in a simple uni-dimensional random sequence was modeled using a MAC model and an exemplar model, the generalized context model (GCM, Nosofsky, 1986).

In this simple implementation of the MAC model, participants are assumed to base their categorization decision for the stimulus on trial n on the difference, d, between the current stimulus, and the stimulus on the preceding trial, trial n-1. Equation 23 uses Gaussian decay to relate the distance d to the probability of responding on trial n with the category label from trial n-1.

$$\text{P(same category)} = e^{-cd^2} \tag{23}$$

The free parameter, c, determines the size of the distance required to give a

change in category label. The Gaussian decay function was chosen because it was a smooth, monotonically decreasing function of d – many other functions would give similar performance. Using Equation 23, the probability of a given response for the last stimulus in any pair of stimuli can be predicted. Note that for some pairs (e.g., 5 followed by 1) the sign of the difference completely determines the categorization. There is no need to rely of the magnitude of the difference. For example, if it is known that category A members take low values on the dimension, and the stimulus on trial n-1 is an A, any stimulus on trial n with a lower value, as indicated by the sign of the difference, must also be a member of category A.

In a truly random sequence every pair of stimuli is equally likely. Therefore by calculating the probability of a correct response for every possible pair, and weighting all these probabilities equally, an average accuracy score can be obtained. The c parameter can then be fit to maximize accuracy. Figure 21 illustrates the predicted probabilities for each stimulus as a function of the preceding stimulus for the category structure illustrated in Figure 20. The predictions shown are for the optimal c parameter which gives an accuracy of 85.2%. For the optimal c parameter the jump size that corresponds to an equal probability of responding with either category is 1.85 tones. However, overall accuracy remains very close to the maximum accuracy for a wide range of c parameters (see Myung & Pitt, 1997). Predictions for stimuli 6 through 10 have been omitted, because, by the symmetry of the category structure, they are the same as the predictions for stimuli 5 through 1. Of interest is categorization accuracy of stimulus 5, which is high when preceded by stimulus 10, but low when preceded by stimulus 1. An exemplar model is unable to predict this pattern of results, as we now show.

The GCM is presented elsewhere (Nosofsky, 1986) but will be described

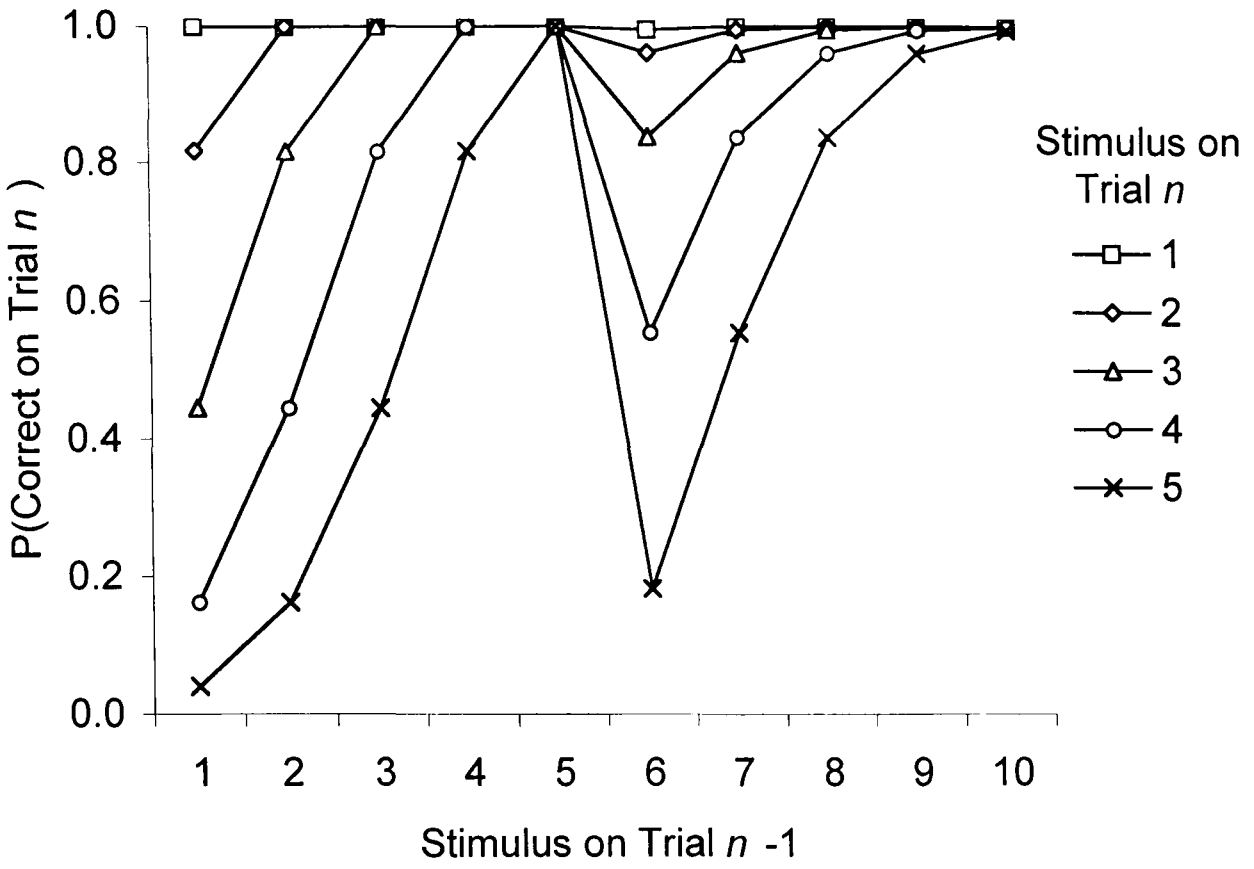


Figure 21. The predictions for the MAC model for the simple category structure illustrated in Figure 21. Accuracy for a stimulus on trial n is plotted as a function of the stimulus on trial n-1.

briefly here. Each stimulus is represented by a vector in multidimensional space (i.e., the stimulus is represented using absolute magnitude information). Each stimulus encountered is stored, together with its category label. The probability by which a stimulus $\underline{x}$ is classified into category $\underline{C_k}$, $\underline{P(C_k|x)}$, is given by the ratio of its summed similarity to examples of that category, $\underline{h_k(x)}$, divided by the summed similarity to all contending categories:

$$P(C_k|\mathbf{x}) = \frac{\beta_k h_k(\mathbf{x})}{\sum_{i=1}^{K} \beta_i h_i(\mathbf{x})} \tag{24}$$

Similarity is monotonically decreasing function of distance, and is typically either an exponential decay or a Gaussian. Thus,

$$h_k(\mathbf{x}) = \sum_{i=1}^{N_k} e^{-cd(\mathbf{x},\mathbf{x}_i)^q} \tag{25}$$

where $\underline{d(x, x_i)}$ is the distance between stimulus $\underline{x}$ and stimulus $\underline{x_i}$ in psychological space, $\underline{q}$ specifies the form of the similarity function, and $\underline{c}$ is a free parameter for the discriminability of the stimuli.

The GCM can be adapted to predict sequence effects by weighting the stimulus on the previous trial more heavily in the summed similarity calculations. In intuitive terms this corresponds to the stimulus either being more available in memory, or being weighted more heavily in the decision process. This means the current stimulus will always be more similar to the category of the preceding stimulus than it would be with no weighting. To demonstrate clear sequence effects, the stimulus on the previous trial was arbitrarily weighted 10 times more heavily than other stimuli. The GCM was used to predict classification accuracies for the category structure described in Figure 20. Figure 22 shows the categorization accuracy for the stimulus on trial $\underline{n}$ as a function of the preceding stimulus on trial $\underline{n}$-

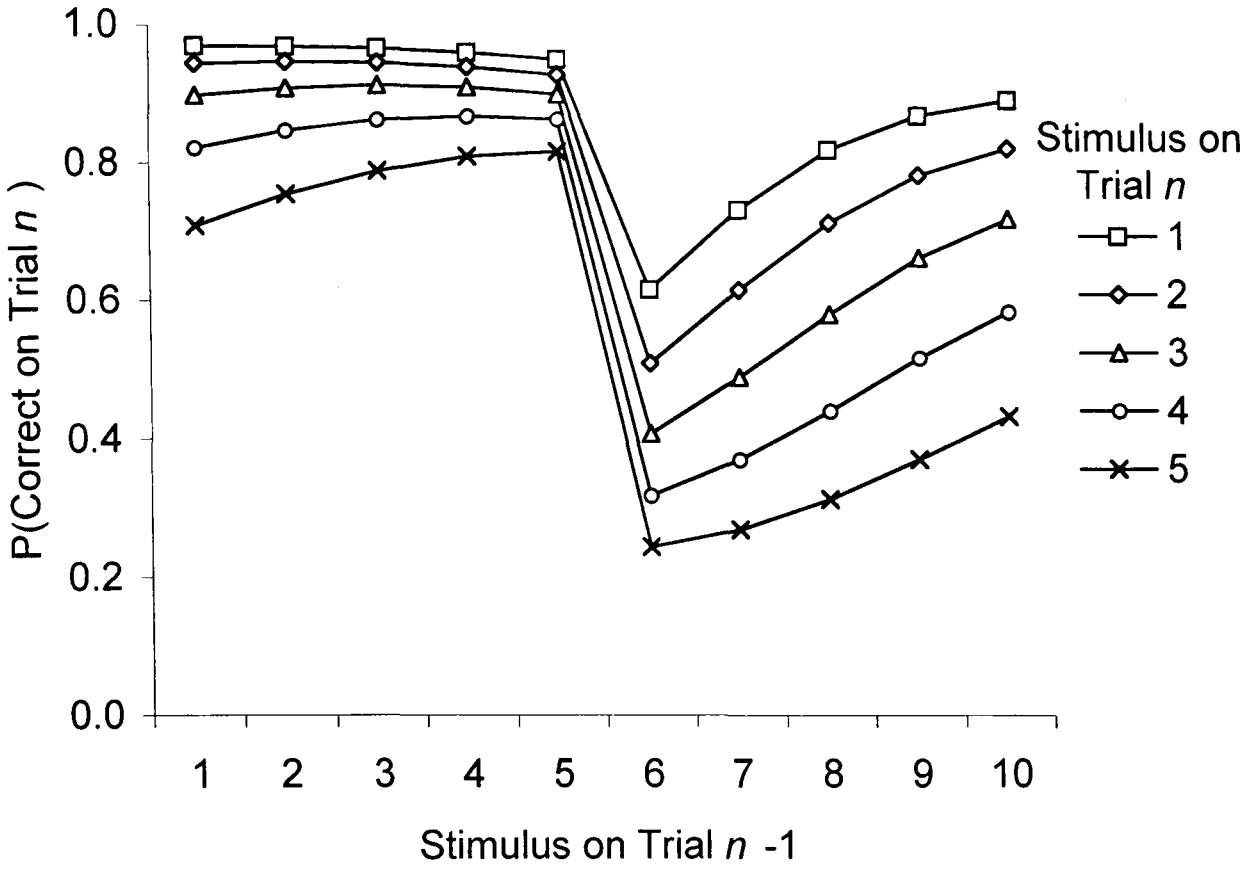


Figure 22. The predictions for the GCM for the simple category structure illustrated in Figure 21. Accuracy for a stimulus on trial n is plotted as a function of the stimulus on trial n-1.

1 for the GCM (q=2, c=0.25, no category bias). Whilst the exact predictions depend on the generalization parameter, c, and the choice of similarity function, q, the qualitative pattern of results is independent of these choices. The optimal value of the c parameter is infinite, as then there will be no generalization between stimuli, and performance will be 100% accurate, with no effect of the previous stimulus. The size of the weighting for the stimulus on trial n-1 also does not affect the pattern qualitatively – a larger weighting simply makes the pattern more extreme. The GCM, unlike the MAC model, is always constrained to predict more accurate classification in the case when the preceding stimulus is from the same category rather than the opposite category.

Overview of Experiments

In the experiments in this paper these opposing predictions for the relative accuracy of classification of a borderline stimulus, preceded by either a distant member of the same category or a distant member of the other category, are tested. All of the experiments use the category structure in Figure 20. The aim was to demonstrate a category contrast effect, whereby classification of borderline stimuli is more accurate when preceded by a distant stimulus from the other category than by a distant stimulus from the same category. A MAC strategy would be able to offer an account of this intuitive potential result, but existing models of categorization would not. The existence of a category contrast effect would therefore provide evidence that categorization is based, at least in part, on relative location information. Experiment 6 uses the frequency of a tone as the dimension of variability in a simple binary classification. Experiment 7 uses simple geometric figures used in categorization experiments where participants have typically been hypothesized to categorize on the basis of absolute magnitude information alone. Experiment 8 is a

replication of Experiment 6, and in addition has blocks of identification, rather than categorization, to allow sequence effects in identification to be measured along with sequence effects in categorization.

Experiment 6

Experiment 6 aims to demonstrate a simple category contrast effect using the category structure in Figure 20. As the concern is with the effect of distant stimuli on the classification of stimuli on the borderline between the two categories, these pairs of stimuli (1 before 5, 10 before 5, 1 before 6 and 10 before 6) are over represented in pseudo random sequences, so that enough data could be gathered in a short experiment. The pseudo random sequences are controlled so that the runs of consecutive categorization responses, the relative frequencies of each tone, and the relative frequencies of each sized jump between tones would be as found in a truly random sequence.

Method

Participants. Ten University of Warwick undergraduates participated in this 10-minute experiment.

Stimuli. Ten 500 ms sine wave tones of differing frequency were used as stimuli in this experiment. Each tone was 1% higher in frequency than the tone immediately lower in frequency, and thus the tones were equally spaced on a log frequency scale. The first tone had a frequency of 600.00 Hz, and the last tone had a frequency of 656.21 Hz. The intention was that adjacent tones anywhere along the scale would be equally discriminable.

Design. The 10 tones were divided into two categories, with the 5 lowest frequency tones in one category, and the 5 highest frequency tones in the other category. Tones were presented sequentially for categorization. Of interest in this

experiment are the effects of the immediately preceding tone (trial n-1) on the categorization of the current tone (trial n). Numbering the tones from 1 (lowest frequency) to 10 (highest frequency), the four critical pairs of tones are 1 before 5, 10 before 5, 10 before 6 and 1 before 6. The pairs 1→5 and 10→6 contain a tone distant in frequency space followed by a borderline member of the same category. The pairs 10→5 and 1→6 contain a distant tone followed by a borderline member of the other category. A simple comparison of the proportion correct on the last trial of each pair for the two pair types (either within category or between category) will allow exemplar and MAC accounts to be distinguished.

Each critical pair was presented once in each block of 20 trials. The four critical pairs were assigned at random to the 4th and 5th, 9th and 10th, 14th and 15th, and 19th and 20th trials in a block. The remaining tones – 2, 3, 4, 7, 8 and 9 – were placed in the unfilled trials at random, subject to the following constraints: (a) each tone occurs equally frequently, (b) the number of occurrences of each size jump in tone reflects the natural distribution of these jumps for a random stream of ten tones, (c) the lengths of runs of tones of the same category is fixed to mimic a random sequence. With these constraints only 42 possible sequences can be generated. For each block a sequence was selected at random from one of the possible sequences. The constraints were designed to allow the critical pairs to be over represented in a sequence without the sequence seeming non-random.

Procedure. Participants were tested one at a time in a quiet room. Participants were instructed that they would hear a number of tones, one after then other. They were told that after each tone they would be asked to respond with one of two labeled keys depending on which category they thought the tone came from. Participants were asked to respond as quickly as possible without making mistakes.

Although at first participants would have to guess, they were informed that by attending to the correct answer displayed on the screen after each response, they could learn which tones belonged to which category. They were given an opportunity to ask the experimenter questions before the experiment began.

Ten blocks of 20 trials were presented to each participant. For each block a different pseudo random sequence, as described in the design, was randomly chosen. Each trial began with a tone, presented for 500 ms, over Sony DR-S3 closed back headphones. Tones were generated by, and response were gathered, using an Apple Macintosh Performa 475. A "?" prompt appeared on the screen with the onset of the tone. From the onset of the tone participants were able to respond with either z or x on a normal qwerty keyboard, labeled "A" and "B" respectively. The assignment of labels to categories was counterbalanced across participants. The "?" prompt disappeared immediately participants responded. After the participants had responded or 1500 ms after the offset of the tone, whichever was later, the correct answer was displayed on the screen for 1000 ms. There was a 500 ms pause before the next trial began. Participants completed all blocks with no breaks between blocks. The experiment took about 10 minutes to complete.

Results

Categorization accuracy reached an asymptote of about 90% correct after the first block of 20 trials. Performance on the last tone in a critical pair is shown as a function of whether the first tone of the pair came from the same category, or the other category (Figure 23). There was a large difference is performance in the two pair types, with participants classifying a borderline tone significantly more accurately after a distant tone from the other category, compared to a distant tone from the same category, $t(9)=3.67$, $p<0.001$. This pattern is the same for both pairs

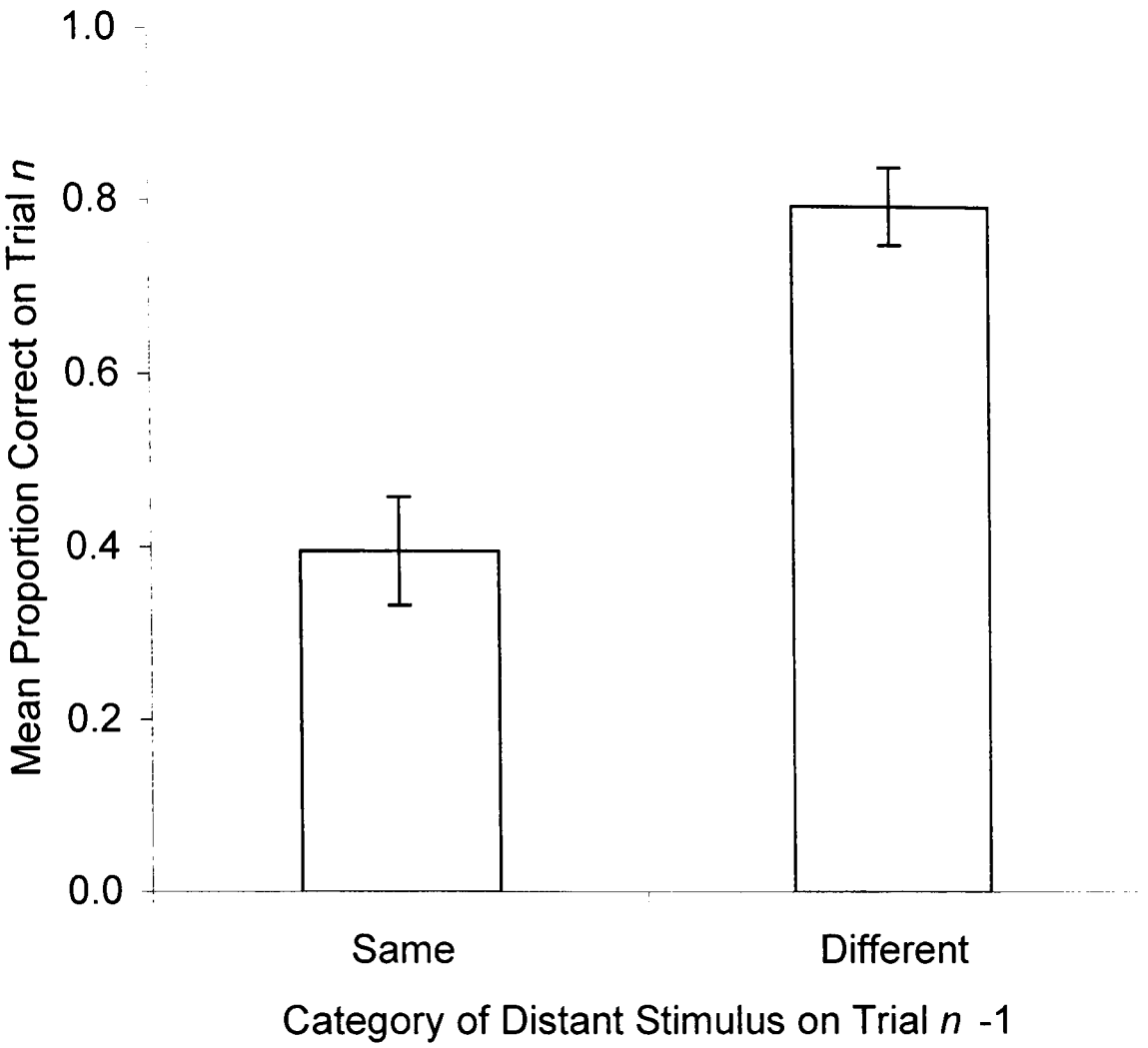


Figure 23. The proportion of correct responses for same category tone pairs (1→5 and 10→6) and different category pairs (1→6 and 10→5) for Experiment 6. (Error bars are standard error of the mean.)

1→5 and 10→5, and for 10→6 and 5→6, and is consistent with a MAC hypothesis.

An alternate explanation of these results needs to be ruled out. The different between the two types of critical pairs is that to get both tones in a within category pair correct participants must make the same category response twice in a row, but to get both tones correct for the between category pair type participants must switch responses. Thus if participants are biased against making two identical response in a row, participants would show poorer accuracy on the final tone of the within category pair than on the final tone of the between category pair. To eliminate this possibility, responses to filler items were examined to measure possible bias. Participants were more likely to persevere with a response than they should be, given the sequence they were presented with. This deviation was not significant, $t(9)=0.47$, $p=0.65$, and is in the wrong direction to explain the pattern of responding on the last item in the critical stimuli pairs.

Discussion

In categorizing a sequence of tones categorization of a tone is influenced by the immediately preceding tone. This finding is consistent with evidence from absolute identification, where there are also strong sequence effects (Lacouture, 1997; Mori, 1989; Mori & Ward, 1995; Ward & Lockhead, 1970; Ward & Lockhead, 1971). When categorizing a tone on the borderline between the two categories, a preceding large within category shift induced significantly more errors than a between category shift (i.e., a category contrast effect is demonstrated). This effect is consistent with a MAC strategy, but not with an exemplar based strategy. Although these effects could potentially be explained by a simple alternation bias (Dember & Richman, 1985) analysis of filler trials in the pilot experiment reveals no evidence of such bias. The category contrast effect is strong evidence that

participants' categorizations are not based on absolute frequency information, but rather on the frequency of a tone relative to the preceding tone(s).

Experiment 7

Experiment 7 is very similar to Experiment 6. The main difference is that the tones were replaced with simple visual stimuli. The stimuli are those used by Nosofsky (1985; 1986) – semicircles that vary in radius, with radial lines that vary in orientation. These stimuli were selected because they are typical of stimuli used in categorization experiments (e.g., Ashby & Gott, 1988; Ashby & Waldron, 1999; Maddox & Ashby, 1993; Nosofsky, 1985; Nosofsky, 1986). Models of categorization applied to data from research with such stimuli assume participants represent the stimuli in a multidimensional space, and therefore make the implicit assumption that participants have access to absolute magnitude information for these stimuli (e.g., the GCM, Nosofsky, 1986; and general recognition theory or decision bound theory, Ashby & Townsend, 1986). If a contrast effect can be demonstrated with these stimuli, then this would challenge this assumption.

Although for the purposes of explication of the MAC strategy we have been assuming that participants do not have absolute magnitude information available to them, we certainly do not claim that this information is completely unavailable. With the simple visual stimuli used in this experiment, it is possible that participants have some absolute magnitude information. Whether this information is directly perceived or deduced from the context the stimuli are presented in is not at issue. However the fact that that the information may be available means that it may be used to inform categorization decisions. If this is the case the category contrast effect is expected to be smaller. Accordingly more participants were tested than in Experiment 6 to detect a potentially smaller effect.

Method

Participants. 26 Warwick University undergraduates and postgraduates participated.

Stimuli. The stimuli used in this experiment were semicircles of varying radius, with radii of varying angle, as used by Nosofsky (1985; 1986). In Nosofsky's experiment, four possible semicircle radii were crossed with four possible radius orientations, to create a 16 possible stimuli, arranged in a four by four grid in diameter-orientation space. In this experiment ten different stimuli were created, arranged in a straight line in diameter-orientation space. Thus both semicircle radius and radius orientation were diagnostic of category. Two alternative spacings of the 10 stimuli were considered. Whilst 10 stimuli spaced equally across a diagonal of Nosofsky's square of stimuli would equate the overall area of stimulus space used by the stimuli, this solution was rejected because the 10 stimuli would be less discriminable from one another than Nosofsky's stimuli, as they fill the stimulus space more densely. It was felt that this would hinder participants in the possible application of an exemplar strategy, as the stimuli would be more confusable. An alternative arrangement (Figure 24) where the 10 stimuli extend outside the region of space occupied by Nosofsky's stimuli was used. Each adjacent pair of stimuli was then spaced as in Nosofsky's experiment. Note that this choice of stimuli is the conservative choice, favoring exemplar models.

Design and Procedure. The design and procedure are the same as Experiment 6, except tones were replaced with a 150 ms presentation of a semicircle with line stimulus in green pixels on a black background.

Results

This analysis is identical to that performed for Experiment 6. As in

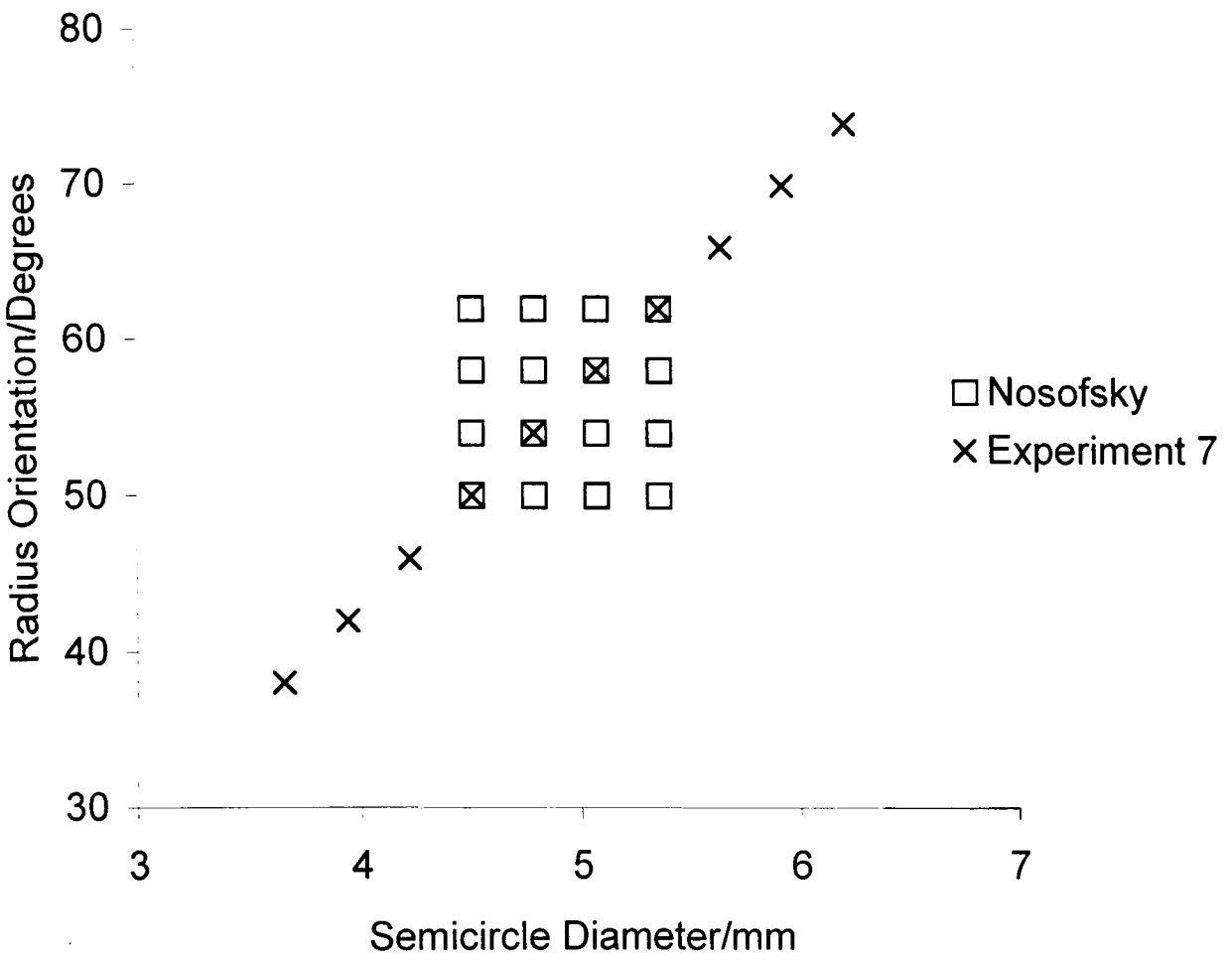


Figure 24. The stimulus structure used in Experiment 7 compared to Nosofsky's (1985) stimulus structure.

Experiment 6 participants quickly reached asymptotic performance of over 90% of filler stimuli correct after one block of trials. Two participants were eliminated from the study for spontaneously reporting that they realized that certain pairs were designed to trick them, and responding to counter this effect. A further participant was eliminated for failing to performance above chance on filler items throughout the experiment. (Note that the filler items were categorized almost perfectly by all other participants.) For the remaining participants, performance on the last semicircle in a critical pair is shown as a function of whether the first semicircle of the pair came from the same category, or the other category (Figure 25). There was a smaller difference in performance in the two pair types than in Experiment 6. The difference however was significant, with participants classifying a borderline stimulus significantly more accurately after a distant semicircle from the other category, compared to a distant semicircle from the same category, $t(22)=3.66$, $p<0.005$. This pattern is the same for both pairs 1→5 and 10→5, and for 10→6 and 5→6, and is consistent with the MAC strategy.

As in Experiment 6, responses to filler items were examined to measure possible bias. Participants were slightly more likely to persevere with a response than they should be. This difference was not significant, $t(22)=1.39$, $p=0.17$, and such a perseverance bias could not explain participants' worse performance in the same condition. (A bias towards giving the same response would reduce errors for these critical pairs, and increase errors for the different category critical pairs.)

Discussion

The category contrast effect demonstrated in Experiment 6 has been replicated in this experiment using different stimuli. The effect was approximately half the size of the effect observed in Experiment 6, consistent with the hypothesis

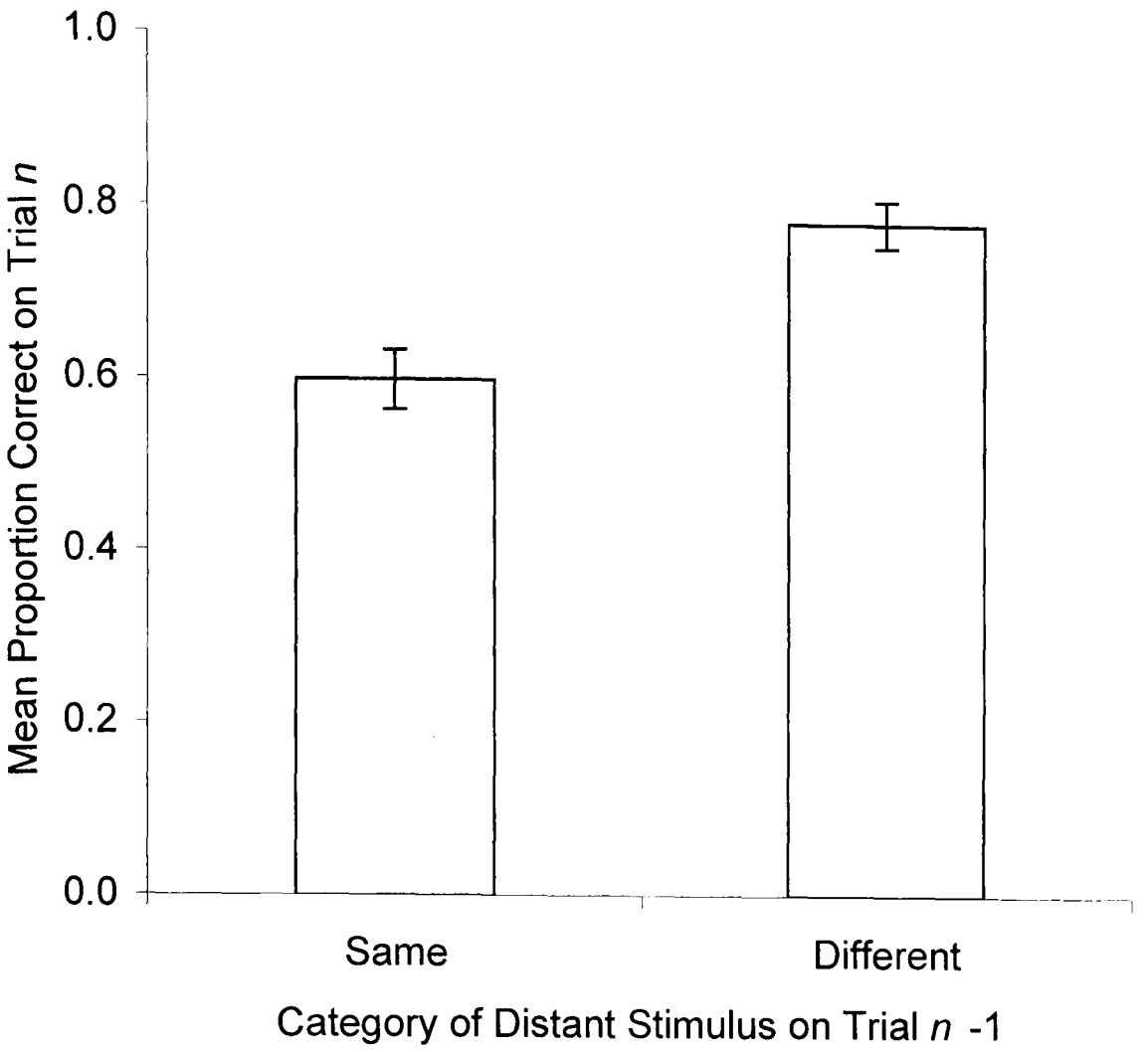


Figure 25. The proportion of correct responses for same category stimulus pairs (1→5 and 10→6) and different category pairs (1→6 and 10→5) for Experiment 7. Error bars are standard error of the mean.

that participants have increased access to absolute magnitude information (but see the General Discussion for an alternative explanation consistent with the MAC explanation). However, the effect is still large, and constitutes a demonstration of a sequence effect in categorization that cannot be accounted for by models of categorization that assume that categorization is based only on absolute magnitude information.

Experiment 8

It is possible that an exemplar model may be able to predict successfully the category contrast effect observed in the first two experiments when sequence effects in absolute magnitude estimation are taken into consideration. The modeling using the exemplar model in the introduction did not take identification assimilation or contrast effects into account. Consideration is given here to the predictions of an exemplar models when assimilation in identification is used as a potential explanation of the sequence effect in categorization demonstrated in Experiments 6 and 7.

In an absolute identification of loudness task the response given to the current stimulus is assimilated to (i.e., correlated with) the immediately preceding stimulus (Lacouture, 1997; Ward & Lockhead, 1970; Ward & Lockhead, 1971). How would identification assimilation affect an exemplar model's predictions for the two types of critical pair of interest? When the distant tone is from the same category, assimilation should cause participants to perceive the tone as more similar to the exemplar of the correct category, and less similar to the exemplar of the incorrect category, than it really is. Identification assimilation will therefore increase categorization accuracy when the preceding tone is from the same category. Assimilation when the distant tone is from the other category will cause participants

to perceive the current tone as more similar the category of the preceding tone, and therefore more similar to the other category, and therefore participants will be more likely to categorize it incorrectly. Therefore identification assimilation causes the exemplar model predicts participants to be even more likely to be correct on a borderline tone when it is preceded by a distant member of the same category, and even less accurate when it is preceded by a distant member of the other category. This effect is in the opposite direction to that needed to allow an exemplar model to explain the pattern of performance observed in Experiments 6 and 7.

However, if with identification of frequency there is an identification contrast effect then an exemplar model would be able to account for the results. There is no evidence for identification contrast with the immediately preceding item in absolute identification of frequency, and indeed such an effect would be at odds with the assimilation observed for other dimensions in previous research. However, such a bias in identification could explain the category contrast effect, without assuming a MAC strategy, as follows. For the within category pair, participants would perceive the borderline tone was further from the distant tone, making it less similar to the correct category that it really should be, and more similar to the incorrect category. For the between category pairs participants would perceive the borderline tone as more similar to the correct category than it really should be, and less similar to the incorrect category. Thus the exemplar model could predict performance to be higher for the different category critical pairs than same category pairs, as observed in Experiments 6 and 7.

Although there is no evidence of contrast to an immediately preceding item in absolute identification, this remains a potential explanation of the results of the first two experiments. Experiment 8 has blocks of identification trials amongst the

categorization trials to simultaneously measure the effect of preceding items in both the categorization and identification tasks. The stimuli are the tone sequences from Experiment 6.

Method

Participants

10 University of Warwick undergraduates took part.

Stimuli and Design

The tones and sequences are the same as those used in Experiment 6. The experiment differs in the addition of identification blocks between categorization blocks. Every two blocks of categorization were followed by two blocks of identification. Identification blocks differed from categorization blocks only in the possible responses available to participants, and the feedback given.

Procedure

The procedure differed from that of Experiment 6 in two ways. First, the initial instructions were changed, to tell participants that sometimes they would do a categorization task, and sometimes they would do an identification task. The instructions for the categorization task were the same as Experiment 6. Participants were told that in the identification task that they should respond with one of ten labels for each tone. They were told that each tone had a unique number label, and that tones were either numbered from highest to lowest (or vice versa, as the assignment of labels was counterbalanced across participants). Participants were told the identification task was quite difficult, and that they should try to get their response number as close as possible to the correct answers.

There were 20 blocks of 20 tones. The first two blocks were categorization, the second two were identification, and so on, alternating every two blocks. Two block runs on each task were chosen to be long enough to allow participants to get a reasonably long run of trials before switching tasks, but short enough to allow several task alternations, to measure categorization and identification sequential

biases as near simultaneously as possible. Before each task switch, brief instructions appeared telling participants to switch tasks. The categorization task trials were the same as in Experiment 6. Identification trials differed only in that participants could make one of ten possible responses using the number keys along the top of a normal qwerty keyboard (with 0 relabeled 10). Feedback was one of the ten labels, displayed in the same way as the feedback category letters in the categorization condition.

Results

Categorization Results. The analysis here follows that used in Experiments 6 and 7. As in Experiments 6 and 7 participants quickly reached asymptotic performance of over 85% of filler stimuli correct after one block of trials. Performance on the last tone in a critical pair is shown as a function of whether the first tone of the pair came from the same category, or the other category (Figure 26). The difference was significant, with participants classifying significantly more accurately after a distant tone from the other category, compared to a distant tone from the same category, $t(9)=3.57$, $p<0.01$. This pattern is the same for both pairs 1→5 and 10→5, and for 10→6 and 5→6, and is consistent with the MAC strategy. The contrast effect is almost identical to that shown in Experiment 6 (Figure 23) demonstrating that the interspersed identification task does not disrupt the effect.

As in Experiments 6 and 7, responses to filler items were examined to measure possible response alternation bias. Participants were slightly less likely to persevere with a response than they should be, showing a very small alternation bias. However, participants did not give significantly more alternation responses than were required for correct performance, $t(9)=1.18$, $p=0.27$. Although an alternation bias could potentially explain participants' performance, the size of the bias here is much too small to explain the large contrast effect.

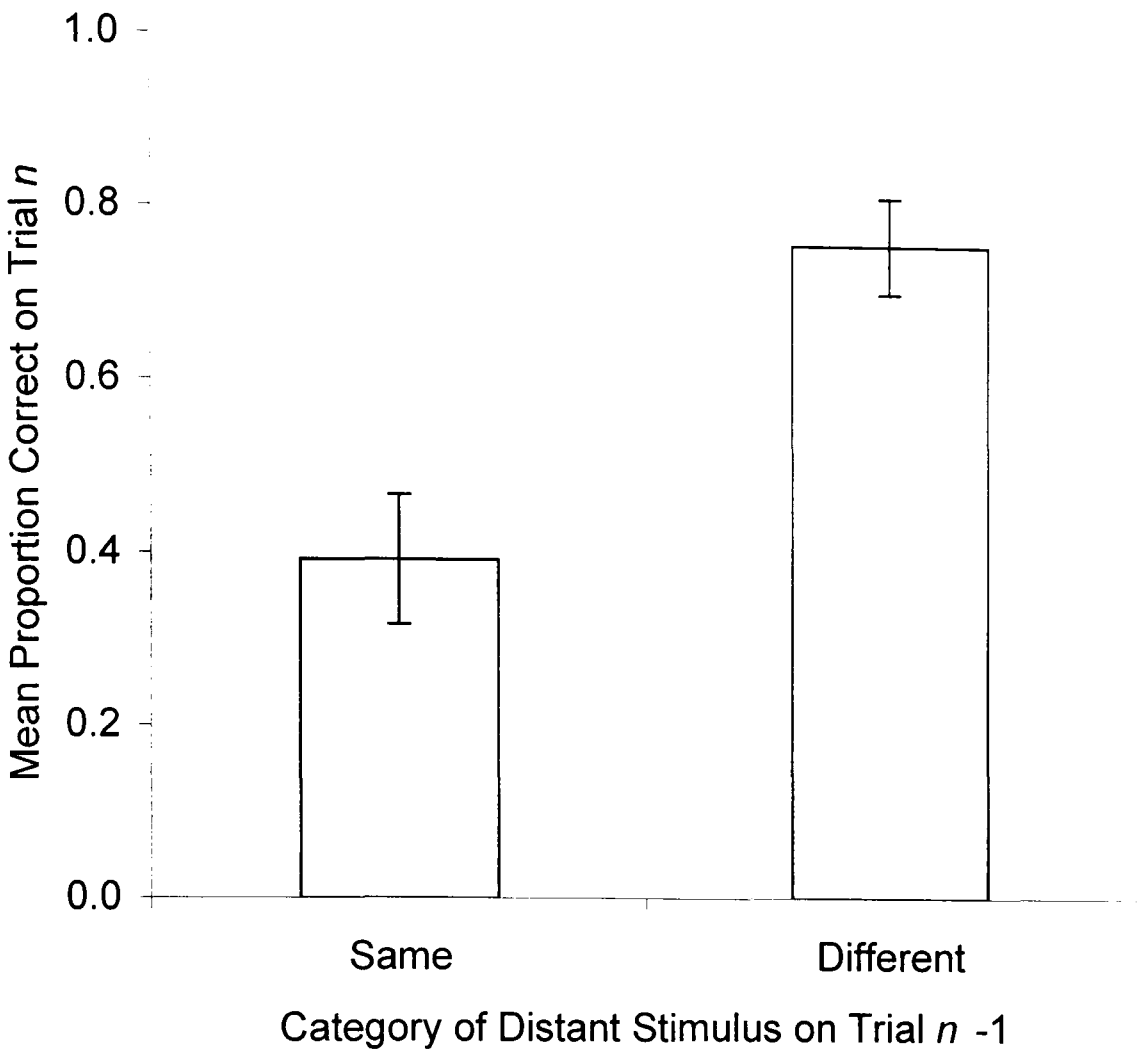


Figure 26. The proportion of correct responses for same category tone pairs (1→5 and 10→6) and different category pairs (1→6 and 10→5) for the categorization task in Experiment 8. (Error bars are standard error of the mean.)

Identification Results. The degree of assimilation was calculated for each of the critical pairs for each participant. For pairs where both tones were from the same category there was an average contrast effect of 0.11 tones across all participants, which does not differ reliably from zero, $t(9)=0.40$, $p>0.05$. For the pairs where the tones were from different categories there was an average assimilation of 0.07 tones, which again does not differ reliably from zero, $t(9)=0.70$, $p>0.05$. The degree of assimilation did not differ between pair types, $t(9)=0.77$, $p>0.05$. The effect of the previous tone on identification of the last tone in a critical pair is very small. Such a small effect is not sufficient to allow an exemplar model to explain the large contrast effect obtained – there would need to be a contrast effect for both pair types approximately one order of magnitude larger. There was reasonable between participants variation in the degree of assimilation averaged across all critical pairs (standard deviation = 0.71 tones), with some participants showing contrast, and others assimilation. The degree of assimilation predicts the size of a participants' category contrast effect, $r^2=0.49$, and this correlation is significant, $t(8)=3.83$, $p<0.01$. Participants showing identification assimilation have a smaller category contrast effect than those showing identification contrast, consistent with the predictions of an exemplar model. However, all participants show a category contrast effect. According to an exemplar model, participants with identification assimilation should show better performance on in categorization on the same category critical pairs than different category pairs, and this is never the case. In summary, whilst the assimilation scores in the identification task are related to the size of the category contrast effect, the assimilation scores do not allow an exemplar model to account for the contrast effect.

Figure 27 shows the average identification error on for a stimulus on trial n

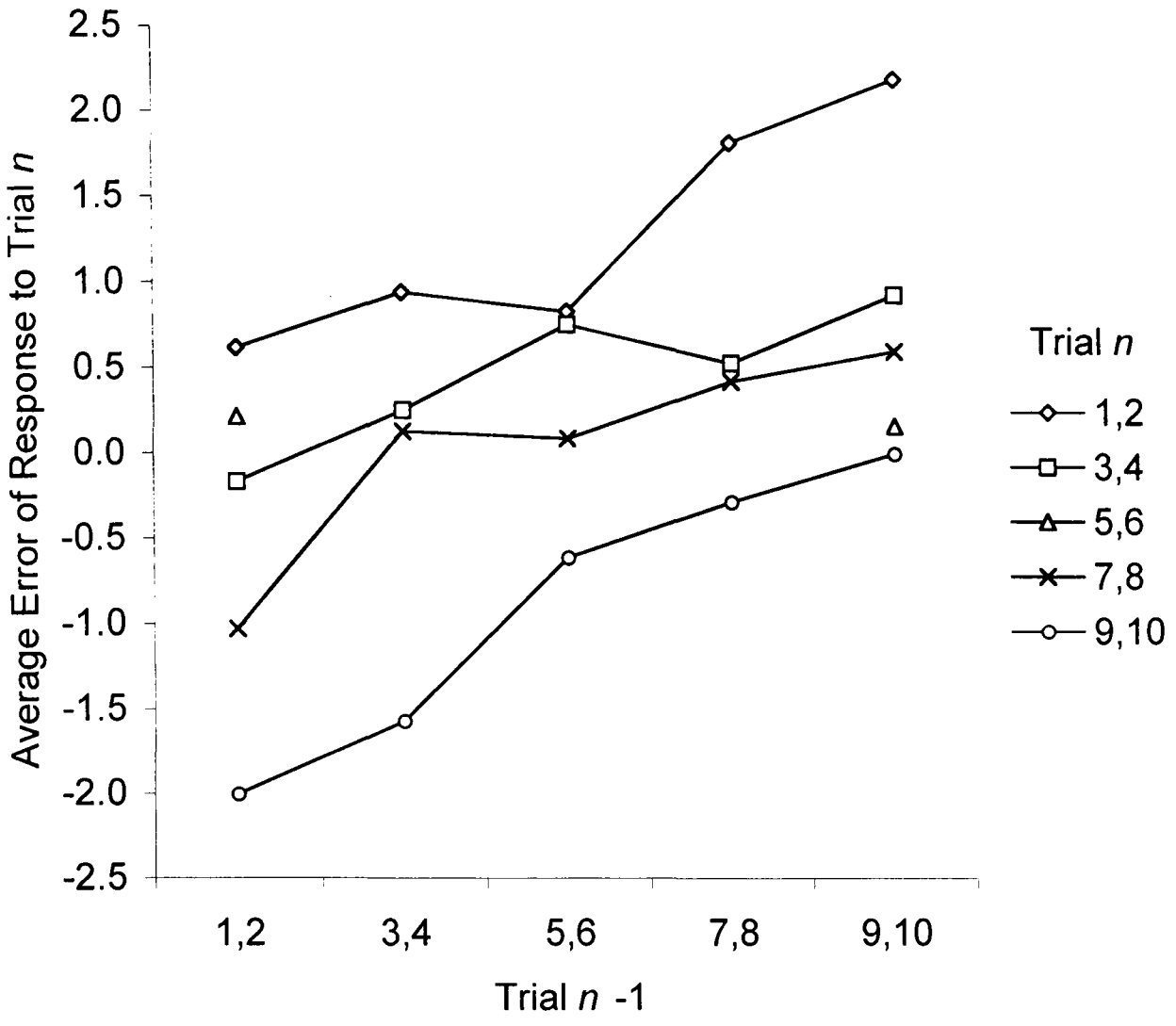


Figure 27. The average error on identification trials plotted as a function of the preceding stimulus for Experiment 8. Adjacent stimuli have been grouped.

as a function of the stimulus on trial n-1, for data from the entire sequence (filler trials and critical trials). This analysis follows Ward and Lockhead (1970). (Tones 5 and 6 only appeared in critical pairs, and were always preceded by either tone 1 or tone 10, and therefore there is no data for tones 5 or 6 preceded by tones 2 through 9.) Data has been collapsed across adjacent pairs of tones to give more trials per data point. Negative error scores correspond to participants underestimating the tones frequency, and positive scores correspond to participants overestimating the tones frequency. Thus the positive slopes of the lines in Figure 27 demonstrates an assimilation effect.

Discussion

The basic category contrast effect from Experiment 6 and 7 has been replicated. Measurement of contrast and assimilation sequence effects in identification in alternate blocks revealed only very small, non-significant effects for the last tones in each critical pair. Thus the possibility that sequence effects in identification may allow an exemplar model to explain the category contrast effect is eliminated. It is worth noting that the category bias parameters in the exemplar model (the β's in Equation 24) do not allow the model to account for the results. Separate bias parameters would be needed for each possible combination of pairs of trials to explain the effect, and thus this explanation offers no explanatory power. Further, in modeling identification, the biases used to predict the categorization data could not predict the identification data.

A relationship between identification bias and the size of the contrast effect was demonstrated. Participants who showed assimilation in identification of the second tone of a critical pair to the first tone showed a smaller category contrast effect than those who showed contrast in identification. Examining the sequence

effects in identification over all identification trials, the standard absolute identification assimilation effect was observed of the current response to the immediately preceding stimulus (Lacouture, 1997; Ward & Lockhead, 1970; Ward & Lockhead, 1971). This the first demonstration of assimilation in identification of frequency of a tone.

General Discussion

A category contrast effect has been demonstrated whereby categorization accuracy of a stimulus near the boundary between two categories is higher when preceded by a distant stimulus from the opposite category than by a distant stimulus form the same category. This large effect persisted throughout each experiment, even after average accuracy reached over 85%. Experiments 6 and 8 demonstrated this effect in a binary classification of tones varying in frequency. Experiment 7 replicated this effect using simple visual stimuli thought by categorization researchers to allow perception of absolute magnitude. We have also found the effect in a meta-analysis of data from other categorization experiments where similar simple geometric figures were used as stimuli (Stewart & Chater, submitted). In these experiments random trial ordering was used, and thus the category contrast effect is not an artifact of the pseudo-random sequences used here. Although the category contrast effects cannot be explained by a simple response alternation bias (Dember & Richman, 1985) as analysis of the responses to filler trials shows no such bias.

The existence of this category contrast effect provides two serious challenges to existing models of categorization. First, the assumption that categorization is based only on absolute magnitude information is challenged, by demonstrating a pattern of errors that can only be accounted for by participants' reliance upon

relative magnitude information. Second, the category contrast effect provides evidence that the local sequential context biases categorization decisions. The category contrast effect is consistent with the MAC account presented here, where in the absence of absolute magnitude information, classification of a stimulus is based on comparison with the preceding stimulus. Experiment 8 explored the possibility that categorization sequence effects may be predicted from identification sequence effects. The tiny non-significant biases in identification were too small to allow an exemplar model to account for the category contrast effect.

The modeling presented demonstrates that exemplar models cannot account for the results. Decision bound models are unable to account for the results when they are adapted to assume the location of the decision bound is altered by preceding material. (Further, the movement of the decision bound seems to relax a major assumption of the model – that absolute magnitude information is available.) To model the category contrast effect, the decision bound would need to move towards the preceding stimulus, so that a borderline stimulus from the same category would fall inside the other category (see, e.g., Treisman, 1985; Treisman & Williams, 1984). This movement of the decision bound would also therefore predict contrast effects in identification, inconsistent with the observed assimilation effects.

Other Sequence Effects in Categorization

Other researchers have investigated sequence effects in categorization. Medin and Bettger (1994) demonstrated that the sequence of training exemplars altered later recognition performance – when training exemplars were sequenced to maximize similarity between adjacent items old/new recognition was improved. Elliott and Anderson (1995) manipulated the order of presentation training exemplars, showing that more distant items were less available for use in a categorization decision, with

the decay following a power law, by item. The number of intervening items was shown to be more important than the intervening time. The concern of these researchers was with the longer term effects of the sequence manipulations, and not with the local sequence effects investigated here.

The Magnitude of the Category Contrast Effect

The sizes of the category contrast effects demonstrated here are smaller than that predicted by the simple MAC strategy outlined here. There are two potential reasons for this: First, participants may be using an improved MAC strategy where comparisons with tones further back in the sequence also inform the categorization decision. Use of this additional information would improve classification accuracy, therefore reducing the size of the category contrast effect. In fact formal modeling has shown inclusion of information from trial n-2 halves the size of the effect. Thus the smaller category contrast effect observed in Experiment 7 may be explained if participants are better able to use information from preceding trials when stimuli are simple geometric figures, rather than tones.

The second possibility is that participants have partial access to absolute magnitude information. In making an absolute identification of length decision, reducing the luminance of lines (with the intention of reducing the amount of absolute magnitude information available) increased the relative contribution information from previous trials compared to information from the current stimulus (Mori, 1989). The idea that reduced availability of absolute magnitude information increases the reliance upon a MAC strategy is consistent with the pattern of the size of effects observed in the experiments presented here. The category contrast effect was larger in Experiments 6 and 8, where tones varying in frequency were used, than in Experiment 7, where simple visual stimuli were used. This is consistent with the

assumption that these simple visual stimuli allow participants more access to absolute magnitude information (either perceived directly, or deduced from comparison with the presentation context).

MAC and Feedback

Many, if not most, perceptual categorization experiments contain blocks where participants are not given trial by trial feedback. (In the experiments presented here participants were given trial by trial feedback.) It would be surprising is category contrast effects were not found in such conditions. Indeed in the absence of feedback very similar absolute identification sequence effects are obtained (Ward & Lockhead, 1970; Ward & Lockhead, 1971). The MAC strategy described here assumes participants have knowledge of the correct categorization of previous stimuli. However adaptation of the strategy to the no feedback conditions is straight forward because of the correlation between the correct answer and the predicted answer. (Even in the simple MAC model presented here, where only information from trial n-1 was used, accuracy was 85%.) A simple solution therefore would be to take the "correct" answer to be that predicted by the model, i.e., A if P(A)>0.5, otherwise B. Alternatively, the response on trial n could be a weighted mixture of the responses calculated for both possible categories of the stimulus on trial n-1, i.e.,

$$P(\mathrm{A}_n) = P(\mathrm{A}_{n-1})e^{-cd^2} + \left[1 - P(\mathrm{A}_{n-1})\right]\left[1 - e^{-cd^2}\right] \tag{26}$$

where $P(A_n)$ is the probability of an A response on trial n, $P(A_{n-1})$ is the probability of an A response on trial n-1, d is the difference between the stimulus on trial n and trial n-1, and c is a free parameter determining the size of the distance required to give a change in category label, as in Equation 23.

Conclusions

A sequence effect in categorization has been demonstrated that challenges the

assumption, implicit in existing models of categorization, that categorization is based on absolute magnitude information. An alternate model has been presented that accounts for this effect by assuming participants instead rely on comparison of a stimulus to immediately preceding stimuli to make a categorization decision.

Chapter 4

Feature Creation in Perceptual Categorization

Abstract

This chapter addresses the claim that features created during experience with novel stimuli qualitatively alter subsequent perception. Schyns and Rodet (1997) provided evidence that categorization creates features, and that these then alter perception. By varying the order of category learning between participants, they were able to induce orthogonal categorizations of identical test exemplars, consistent with the hypothesis that participants learned different sets of features. Experiment 9 provides a replication of this feature creation effect. Experiment 10 manipulates the presentation context of the test exemplars. The absence of a feature creation effect in this experiment suggests that the categorization of the exemplars is dependent on participants' beliefs that a compound stimulus is made from a conjunction of other stimuli. Using simple, random line drawings, Experiment 11 shows variable categorization of test stimuli analogous to those of Schyns and Rodet that is not driven by either order of category learning or beliefs that a compound stimulus is made from a conjunction of stimuli. A new feature creation paradigm is developed using black and white checkerboard stimuli. Invariant patches of squares embedded in random checkerboards are diagnostic of the checkerboard category. Experiments 12 and 13 use a design similar to that of Schyns and Rodet, and together provide evidence for feature creation. Experiments 14 to 16 provide solutions to three methodological problems in using Schyns and Rodet's design with checkerboard stimuli: (a) the avoidance of sequential presentation of parts of the exemplars on test allows reaction time measurement; (b) the need for participants to remember features over long blocks of intermediate trials is eliminated; (c) a confound between possible feature location and feature creation effects is eliminated. A feature creation effect is demonstrated that is consistent with participants learning the largest invariant part of

a checkerboard as a single feature (Experiment 16). Further, it is shown that learning of a large configuration of squares can be blocked by prior experience with parts of the configuration.

Feature Creation in Perceptual Categorization

A feature refers to "any elementary property of a distal stimulus that is an element of cognition", or "an atom of psychological processing" (Schyns, Goldstone, & Thibaut, 1998). This chapter investigates the possibility that new features may be created to serve new categorizations, and further that these features may qualitatively change the perception of stimuli. Features are identified by their role in cognition: for example, they allow new categorizations and perceptions to occur. (Throughout this chapter the word feature is reserved for the representation of part of a stimulus. The word element describes a physical part of a stimulus, and the word type describes an entire category of stimuli.) Why might the creation of new features be important? The amount of data needed to specify a mapping between possible stimuli (input vectors) and categories (output vectors) increases exponentially with the number of dimensions of the input vector (the curse of dimensionality, Bellman, 1961). Everyday stimuli typically exist in a space with a large number of dimensions, e.g., approximately 35 dimensions for faces (Hinton, 2000). Two helpful observations can be made: (a) Often the assignment of each output vector to each input vector is not arbitrarily. Thus if data are absent for a region of input space it is possible to interpolate between near by data points. (b) Most real data sets do not fill the entire possible input space, but instead occupy some lower dimensional subspace. Recoding the data in a lower dimensional space allows training data to fill better the input space. Often a recoding allows a category learning mechanism (supervised or otherwise) to learn better the relevant mapping (Bishop, 1995), provided, that is, the improvement in learning caused by the reduced dimensional space is not offset by information loss in the reduction. The task is to find the dimensions of this subspace. Features, then, allow a high dimensional space, sparsely

filled by a data set, to be reduced in dimensionality, so that the set now better fills a lower dimensional space. (Note that the recoding of space using the new features, causing a reduction in dimensionality, is different to the features simply augmenting the perceptual space, when dimensionality would be increased.)

Evidence for the Creation of New Features

Empirical work supporting the hypothesis that features can be created is found in both the categorization and the attention literatures, and is briefly reviewed here. The idea that experience can create new features is certainly consistent with the large body of evidence showing that such experience alters perception (perceptual learning – for an extensive review see Goldstone, 1998).

Shiffrin and Lightfoot (1997) have demonstrated feature learning effects during visual search of alphanumeric-like characters. Participants were given the task of finding a target that differs from distractors by a unique combination of line orientations at particular locations. With practice the decision latency cost of additional distractor stimuli in a search for conjunctions of features was hugely reduced from hundreds of milliseconds per stimulus to only tens. (Although a small slope still remained after practice, it is possible for a parallel search process to predict such a small slope, if the time taken for a detector in each location to report in parallel is noisy.) When participants were transferred to a new search with new characters that did share individual line segments with the old characters, but that did not contain any of the same combinations of line segments with the old character set (i.e., there were no common configural features), there was no transfer of the reduction in search slopes. Therefore, reduced search slopes could not be explained by general experience with the class of stimuli. When targets and distractors were re-paired, the reduction in search slopes remained. Whatever was responsible for the

reduction in search slope, it was not specific to particular target-distractor pairings in particular displays. Experimentally reducing the within-character set similarity did not lead to a further reduction in search slopes, suggesting that it was not a general reduction in stimulus similarity with experience that was important. Whatever was learned in visual search, it was specific to particular targets, and specific to particular distractors – the results are certainly consistent with a feature creation hypothesis, whereby line segments of the characters are unitized to form a single feature. This result stands in contrast to Triesman and Gelade's (1980) demonstration that extensive practice on a conjunction of color and shape search task did not significantly reduce search slopes.

Further experiments suggest that although Shiffrin and Lightfoot did not manage completely to eliminate the effect of additional distractor stimuli with practice, new features can be processed in parallel across the visual field – which is traditionally taken as evidence for the existence of a feature (Duncan & Humphreys, 1992; Duncan & Humphreys, 1989; Treisman & Gelade, 1980; Treisman & Sato, 1990; Wolfe, 1994; Wolfe, Cave, & Franzel, 1989). Shiffrin and Schneider (1977) trained participants to asymptote on a varied mapping (VM) visual search, where the target was selected at random from the stimulus set on each trial. The remaining characters in the set were used as distractors. Thus a target on one trial may have been a distractor on the next, and a distractor on another trial may have been a target on the next. A further reduction in search latency was observed when participants were switched to a consistent mapping (CM) search. (In a CM search the stimulus set is split into targets and distractors. Target stimuli always appear as targets, and distractor stimuli always appear as distractors.) Shiffrin and Schneider suggest that this was because of the development of automatic attraction of attention to the CM

stimulus. Such a finding is further supported by the fact that an old CM target will draw attention away from the current target in a visual search, even though participants are actively told to ignore the stimulus in the location the old target appears in (Shiffrin & Schneider, 1977). It seems then that single features may be formed for targets in visual search, by unitizing the elements of the stimuli. If CM targets do indeed come to attract attention to themselves, search for a CM target should not interfere with search for a VM target, provided that the two targets do not occur on the same trial. Schneider and Fisk (1982) gave participants two simultaneous visual search tasks. One search was a CM search, and the other was a VM search. In both tasks participants searched for target digits amongst distractor letters. In an initial experiment, the two were found to interfere with one another, suggesting that even after extensive practice, CM search was not automatic. However, when participants were instructed to concentrate on the VM search only, and only to respond to a CM target if they happened to notice it, participants performed as well on both tasks simultaneously as they did on either singly. There was no evidence of dual task decrement at all. It seems that a CM search can become automatic even if search slopes will not become flat (although search slopes were not measured in this study).

There is also evidence that distractors in visual search may also be unitized, as they are demonstrated to develop attentional properties of their own. After CM visual search training Dumais (1979; Shiffrin & Dumais, 1981) transferred participants to several transfer conditions. Targets were left unaltered and retrained with new distractors, either from a VM search, or novel stimuli. Alternatively, distractors were retrained and targets replaced by either stimuli from VM search or novel stimuli. Almost complete transfer was observed in all conditions. Of interest

here is the large transfer when target stimuli were replaced with familiar stimuli from a VM search, and distractors remain the same. That transfer occurred cannot be explained by familiarity differences. The finding is consistent with the explanation that distractors in the CM search have reduced ability to attract attention. Conditions with novel stimuli differed little from conditions with VM stimuli, suggesting VM training does not significantly alter stimulus's ability to attract attention.

The evidence reviewed thus far suggests that features may be created in visual search. The similarity between categorization tasks, where participants see a stimulus, and respond on the basis of the diagnostic features amongst non-diagnostic features, and visual search, where participants respond to the presence or absence of a target amongst distractors, suggests that categorization experience will also lead to the creation of features.

Goldstone (2000) provided evidence of unitization of diagnostic features in a categorization task. In the task participants classified a stimulus made from five complex, curved line segments. The classification could only be made by attending to all of the five segments, no single segment or two, three or four way conjunction was sufficient. Goldstone hypothesized that participants were able to unitize their representations of the five line segments to create a new, single diagnostic feature, when the segments always appeared in the same order. Compared to a task where the segments did not appear in the same order, removing the possibility of creating a single unitized feature, there was a larger reduction in categorization latency with practice for the stimulus with consistently ordered parts. In a categorization where one segment alone was diagnostic, the practice effect was much smaller. Using data from this single segment diagnostic task, latency predictions were generated for classification models where the decision is based on integrated evidence from the

detection of the five line segments separately. Some classifications of the consistently ordered stimulus when only the conjunction of all five segments was diagnostic were shown to be reliably faster than could be predicted by these models, even when they were granted with unlimited, parallel processing of each of the five segments. A model where the detection of segments interacts, so that the detection of one facilitates another, could account for these data, but allowing such interaction is surely equivalent to at least partial unitization of segments. This is because the facilitation would have to be specific to particular orders of line segments to explain the difference between the consistent and randomly ordered conditions. A number of boundary conditions provide useful information on the limits of the unitization observed. Goldstone showed that even when the line segments were physically disconnected, slightly smaller but still substantial practice effects were obtained, suggesting that non-connected segments may still be unitized. If the stimuli were made so large that they could not be viewed clearly without a saccade, the practice effect was greatly diminished.

It is also possible that non-diagnostic features in a categorization task could be unitized. Preliminary support for this idea comes from a checkerboard categorization experiment by Graham and McLaren (1998). They demonstrated that pre-exposure to non-diagnostic features in a categorization task retarded subsequent discrimination between those features. In order for subsequent discrimination learning to be impaired, relative to unexposed control checkerboards, something specific to the non-diagnostic exposed stimuli must have been learned. Graham and McLaren account for their results in terms of negative priming (Tipper, 1985) of the learned representations of the distractor stimuli.

Further suggestive evidence that features may be formed during experience

with complex stimuli comes from the face inversion effect (Yin, 1969). The face inversion effect is that faces are recognized worse when upside down, and further that the deficit caused by inversion is larger than for other control stimuli such as houses (Yin, 1969) or landscapes (Diamond & Carey, 1986). Diamond and Carey (1986) suggest the face inversion effect is caused by familiarity or experience with the stimuli, and demonstrated that gun dog experts show a larger inversion deficit in gun dog recognition than do non-experts. These effects are also seen with novel stimuli exposed in an experimental setting. McLaren (1997) investigated a checkerboard analogue of the face inversion (see Gauthier & Tarr, 1998 for a similar experiment). He trained participants on checkerboards, and showed that the inversion hindered subsequent discrimination learning and recognition of the familiar checkerboards compared to novel checkerboard controls. Further, the effect was contingent upon the learned categories having a prototypical structure, as demonstrated by the absence of an inversion effect when the learned exemplars were generated in such a way that the central tendency of each category was not a checkerboard, but a set of gray columns. Thus it seems that during categorization what is learned, and what is subsequently destroyed on inversion, is the prototypical structure of each category. Palmeri and Nosofsky (in press) show that after experience with checkerboard categories very similar to those used by McLaren, the prototypes of the categories are in fact extreme points in psychological space, rather than the central tendency of the category. If the checkerboards are indeed classified using new learned representations of the prototypes, i.e., new features, then this is exactly the rearrangement of perceptual space that would be expected.

Evidence that New Features Qualitatively Change Perception

The evidence reviewed thus far suggests that categorization experience

creates features. This conclusion is certainly consistent with the changes in perception that occur after categorization experience. A pair of stimuli that falls across a category boundary may become more discriminable than an equivalent pair that falls within a category, when both pairs were equally discriminable before learning of the categorization (Goldstone, 1994), although learning that stimuli belong in the same category does not reduce their discriminability (McLaren, Leevers, & Mackintosh, 1994). This chapter is concerned with a further question. Can new features qualitatively change the perception of stimuli, rather than simply facilitating the processing of stimuli? Indirect evidence that this may be the case is provided by Goldstone (1995). Participants were asked to adjust the color of an initially black simple shape to match the color of another, simultaneously displayed reference copy of that simple shape. The reference copy of each shape always appeared in the same color. Each shape belonged to one of two shape categories. Although the shapes of the objects were irrelevant to the participants' task, Goldstone found that color matching was influenced by the shape category of the object. For shapes of the same reference color, but different shape category, the color of the adjusted shape was systematically set nearer to the average color of the shape's own category. A second, similar experiment provided data consistent with interpreting this bias as the cumulative result of with-in category assimilation and between category contrast. Thus the perception of a feature has been shown to be altered by categorization experience. However, this effect can be explained without hypothesizing the creation of new features, for example, by assuming that participants are sensitive to the correlation between two features.

More direct evidence that new function features create a change in the perception of the stimuli is provided by Schyns and his colleagues (see Schyns et al.,

1998 for a review). Their studies may be divided on the basis of the dependent measure used. Some studies used participants' categorization responses, but first studies based on delineation of parts of stimuli will be discussed.

Schyns and Murphy (1991, and see also Schyns & Rodet, 1997 for a related experiment) taught participants about novel "Martian rock" stimuli. The rocks were 3D gray shaded objects displayed and rotated on a computer screen. The rocks had protrusions and dents all over their surface, making them look very irregular. In the first phase of the experiment participants saw examples of type A rocks. The rocks all had a common protrusion, element a, and after experience with the rocks participants circled or delineated element a as the important invariant part of the rock. (Naïve participants showed little consistency in their decompositions, and did not delineate element a – initially the random protrusions were more salient than the target features.) In a second phase of the experiment, participants learned a new type of rock, type AB, with a new invariant protrusion, element b, adjacent to the previous element, element a, forming a new invariant protrusion, element ab. After experience with the new rocks, participants delineated both element a and element b separately. Relevant here is the contrast with naïve participants, who were not exposed to the type A rocks. They also learned to delineate the invariant part of the new rock, but delineated it as a single part, (i.e., delineated element a and element b as a single feature). Participants' delineations were altered by experience with a category. Their segmentations of rocks differed from that before training despite a strong bottom up constraint – the minima rule (Hoffman & Richards, 1984). Further, the features learned for one category of Martian rock affect the features for subsequently learned categories.

Schyns and Murphy (1994) demonstrated that after participants had learned a

rock feature, the feature could be broken down to allow separate delineation of parts with further categorization experience. In phase one of the experiment, participants learned type AB boulders, and delineated element a and element b together as a single part. In phase two participants learned either type A or type B boulders, and delineated either element a or element b alone accordingly. In phase three participants were given more type AB boulders to parse. They now delineated element a and element b separately. They divided into two parts that which previous delineations indicated was perceived before as a coherent whole.

The extent to which these experiments may be taken to support a feature creation hypothesis depends on the assumption that parts delineated correspond to the features participants gain from exposure. Participants may have been circling parts they felt were diagnostic, rather than parts they had previously seen as they were instructed to do. Certainly delineation of parts is not free from cognitive influence (Schyns & Murphy, 1994). The use of delineation implies that participants have conscious access to the parts they are using, and that their choice of delineations is not influenced by biases or preconceptions of what kind of responses the experimenter desires. The use of categorization as an alternate measure of the hypothesized learned feature set goes some way towards addressing these criticisms.

In experiments by Schyns and Rodet (1997) participants learned to categorize different types of stimuli one type after the other. The learning of features for the early types is designed to alter the features learned for later types, creating different feature sets for different category learning orders. If different feature sets are indeed learned then it should be possible to get mutually exclusive categorizations of identical test stimuli as a function of training category learning order. In other words, identical test stimuli are classified into different categories, suggesting that different

participants perceive the same stimulus differently.

The following experiments are described in some detail as the design and stimuli used provided the basis for the experiments in this chapter. In Schyns and Rodet's (1997) Experiment 1, two categories of "Martian cells" were learned. Martian cells are stimuli that look like normal cells viewed under a microscope. Each cell consisted of a series of random, dark cell bodies inside a gray circular cell. Type A cells contained a particular shaped cell body element a and other random cell bodies. Type AB were cells containing an element made from the conjunction of element a and element b, a different shaped cell body, as well as other random cell bodies. (In fact, the features were counterbalanced, so that some participants learned type B in place of type A.) Half of the participants learned the types in the order A→AB and the other half learned the types in the reverse order, AB→A. The A→AB group should learn the feature set {a, b}, and the AB→A group should learn the feature {ab, a}. In the test phase of the experiment four types of cells were presented. Instead of seeing the entire cell, participants were told they would see two close-up images of parts of the same cell, one after the other. They were instructed to respond only after seeing both close-ups. For type A snapshots, one snapshot was of element a, and the other was a distractor feature. For type AB snapshots, one snapshot was element a and element b together, and the other was a distractor. For type A-B snapshots, one snapshot was element a, and the other snapshot was element b. For distractor cells, both snapshots were distractor features. Participants were asked to classify these snapshots as belonging to either the first or second category they learned. The result of interest is the classification of the A-B snapshots. Although snapshot A-B contains element a and element b, the elements do not appear together. Thus group A→AB, who identify type AB by the co-occurrence of

two features, a and b, should classify snapshot A-B as type AB, as both feature a and feature b will be present. However, group AB→A, who identify type AB by the presence of a single feature, ab, should not classify snapshot A-B as type AB, as feature ab is not present. Group AB→A classified the snapshot as type A the majority of the time, but group A→AB classified the snapshot as type A significantly less often. The group AB→A classification of type A-B snapshot as type A was consistent with the hypothesis that they do not have feature b, that is, they do not have a cognitive representation element b. Was group AB→A really blind to element b? In a final stage participants delineated type A, type B and distractor cells. Unlike group A→AB, participants in group ab→b did not delineate element b.

As Schyns and Rodet (1997) pointed out, there is an alternate attention based explanation of the results. Assume all participants use the fixed feature set {a, b}. Group A→AB participants may assign more attention to feature b, as it is particularly important in discriminating between type A and type AB cells. Less attention may have been assigned to feature b in the group AB→A, as without prior knowledge of the discrimination it is not obvious that feature b is the critical discriminating feature. Thus the differing categorization of the type A-B snapshot pair could just be due to changes in selective attention. A second experiment rules out this possibility. Here two groups of participants learned three categories, one after the other: group A→B→AB, and group AB→A→B. This ensures that both feature a and feature b are equally diagnostic, and therefore equal attention should be paid to each, eliminating the possible selective attention account of the orthogonal categorization of the critical snapshot A-B. Categorization of the type A-B snapshot was consistent with group A→B→AB having feature set {a, b}, and group

AB→A→B using {ab, a, b}. Group A→B→AB almost always categorized the snapshot as type AB, and group AB→A→B categorized the snapshot as type A or type B, but only rarely as type AB. In summary, Schyns and Rodet's (1997) experiments provide evidence that features are created, and that these features alter qualitatively the perception of a stimulus, so that participants with different feature sets give different classifications of this stimulus.

Overview of Experiments

The experiments described thus far provide compelling evidence that new features may be created. The experiments in this chapter are concerned with a further claim that learning these features qualitatively changes the perception of stimuli, rather than simply speeding the processing stimuli. That is, if participants learn new features, they will not only process the stimuli more efficiently, but the perceptual space in which they represent the stimuli will be reorganized. To investigate this, categorization is used as the dependent measure in all of the experiments in this chapter, as in the studies of Schyns and Rodet (1997). If participants' perceptual spaces have been reorganized, then it should be possible to get participants with different feature sets to classify identical stimuli differently. The first two experiments in this chapter are replications of Schyns and Rodet's (1997) Experiment 2. Experiment 9 replicates their feature creation effect, but Experiment 10 shows the effect can be eliminated by adding a background context to the snapshots presented in transfer. This result suggests that the categorization of the snapshots depends on participants' beliefs about whether the elements shown in snapshots may be joined onto other elements. Experiment 11 represents an attempt to establish a feature creation effect in a new line drawing stimuli paradigm, and to see whether participants' awareness that the compound element is made up of the two

other elements is a good predictor of their performance on snapshot stimuli. Experiments 12 and 13 have a similar design to Schyns and Rodet's experiment, but use black and white checkerboard stimuli. By using these new stimuli, the need for sequential presentation of parts in the test phase is eliminated. However, participants struggle to maintain the memory of the features learned at the beginning of the experiment throughout the experiment. Experiment 14 demonstrates a feature creation effect using a new paradigm that avoids the need for participants to maintain features in memory. Experiment 15 demonstrates a possible feature location confound may be available as an alternate explanation in Experiment 14, but Experiment 16 demonstrates a feature creation effect that cannot be explained by such a confound.

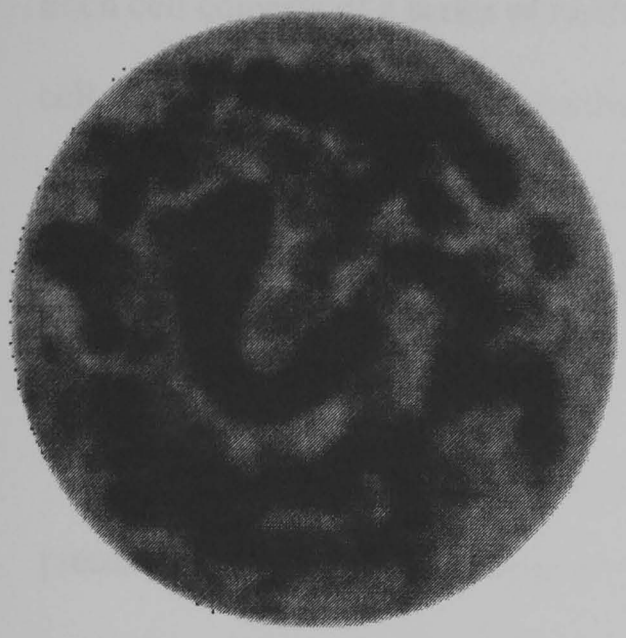
Experiment 9

This first experiment is a replication of Schyns and Rodet's (1997) Experiment 2.

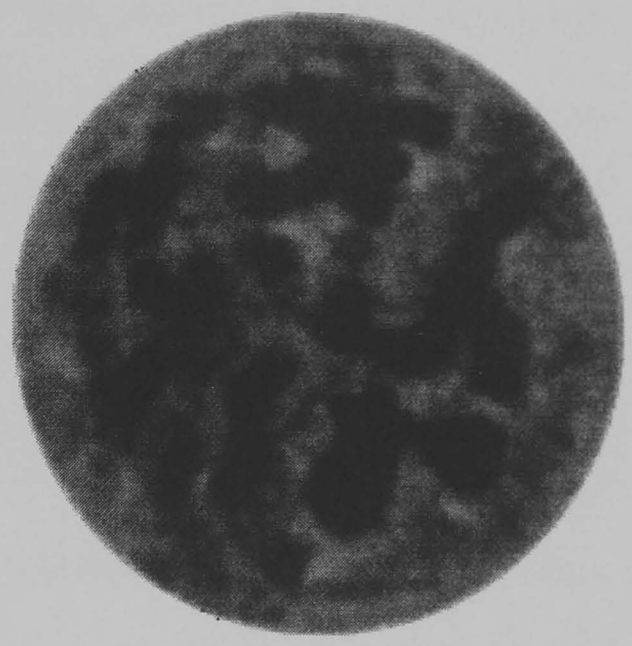
Method

Participants. 32 undergraduates from the University of Warwick participated, either to receive course credit or monetary payment. The experiment was run in a one hour session with other unrelated experiments involving checkerboard stimuli (not those experiments presented in this chapter). Participants were randomly assigned to one of the two experimental conditions.

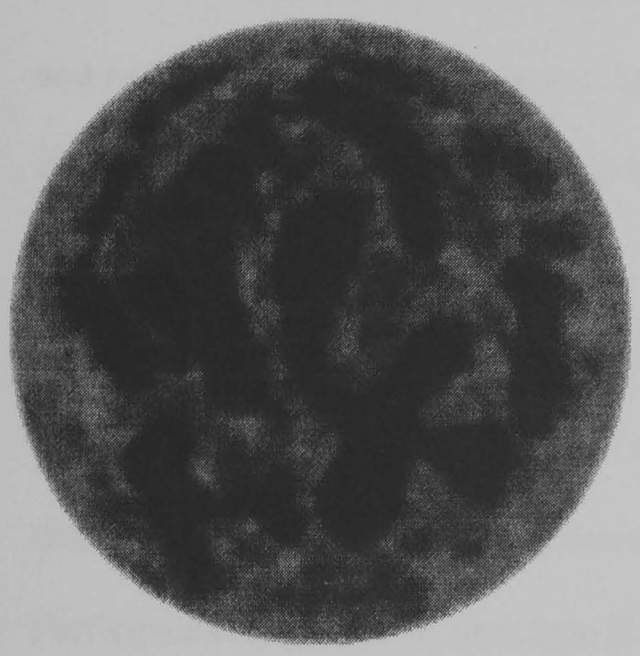
Stimuli. The stimuli used were the same as those used by Schyns and Rodet (1997). Two kinds of stimuli were used in this experiment. During the learning and verification stages entire Martian cells were presented for 3 s (Figure 28). (In Schyns and Rodet's (1997) original Experiment 2, cells were presented for only 2 s. The increase in presentation time here was designed to facilitate learning of the cells.)



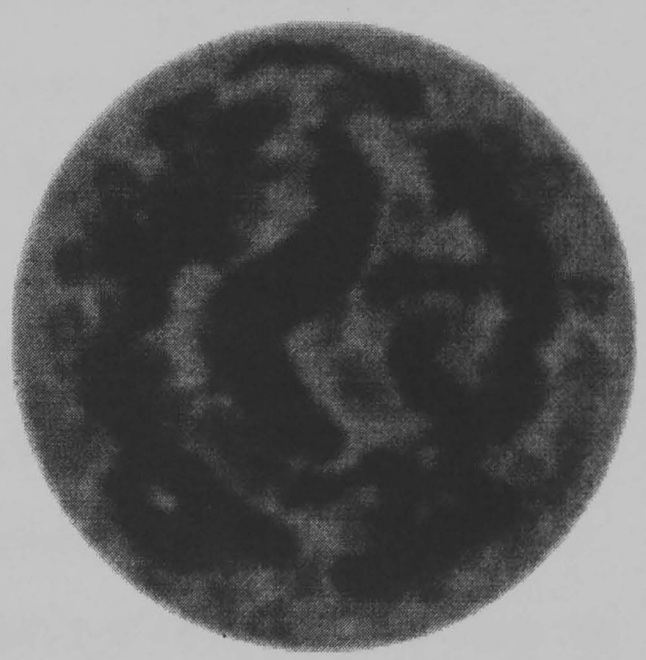
Type A



Type B



Type AB



Distractor

Figure 28. The Martian cell stimuli used in the learning and verification phases of Experiment 9.

Each cell consists of a series of random cell bodies. There are four types of Martian cell: type A, type B, type AB and distractor cells. Type A cells contain element a as one of the cell bodies. Type B cells contain element b. Type AB cells contain a compound element made from joining elements a and b together. Distractor cells contain only random cell bodies.

During the final test phase, two close-up snapshots from a Martian cell were presented, each snapshot lasting two seconds (Figure 29). There were four kinds of snapshot pairs: A, B, AB and A-B. Snapshot pair A was a close-up of element a and a close-up of a distractor body. Pair B was a close-up of element b and a close-up of a distractor body. Snapshot pair AB contained a close-up of compound element ab and a close-up of a distractor body. Snapshot pair A-B contained a close-up of element a and a close-up of element b (where the - denotes stimuli appearing separately).

Design. There were three phases in this experiment, a learning phase, a verification phase and a test phase. The order in which participants learn about the three types of Martian cell is a between participants factor with two levels. Participants learn type A, then type B and finally type AB (A→B→AB) or type AB, then type A, then type B (AB→A→B). The order of learning types a and b was counterbalanced. It was hypothesized that participants in condition A→B→AB would have feature set {a, b}, but that those in condition AB→A→B would have an additional feature in their set {ab, a, b}. This is because in condition AB→A→B a feature (feature ab) must be created to represent the ab element encountered first, but that in condition A→B→AB features a and b can be used together to represent the element ab.

Procedure. The experiment took place in a quiet room. Participants were

Snapshot of Element a

Snapshot of Element b

Snapshot of Element ab

Snapshot of Distractor Part

Figure 29. The Martian cell snapshot stimuli used in the transfer phase of Experiment 9.

seated in front of the computer and the keyboard and monitor were adjusted as necessary. Participants read the instructions on the computer screen and were given an opportunity to ask the experimenter questions. They were instructed that they would learn about types of Martian Cell. They were told that cells belonging to the same type had the same diagnostic cell body in them, and that to learn the types they should search for cell bodies common between cells of the same type. The experiment then started with the learning phase. Participants were instructed that they would see ten cells from the same type, and that they should try to spot the cell body common to that type. There were ten learning trials. Each trial was preceded by 2 s of blank screen, followed by a 3 s presentation of an example of a "type 1" cell. After the last trial the learning verification stage began. Four cells were displayed in a random order, two new examples of "type 1" cells and two distractor cells. There was a 2 s blank before each cell, a 3 s presentation of the cell, a 0.5 s blank and then response prompt text (e.g., "type 1 or other") which remained on screen until the participant had responded. Participants were instructed to categorize the cells as "type 1" or to press "other" if they thought the cell was not type 1. Participants then responded by pressing keys labeled "type 1" or "other". Keys z, x, c and v on a normal qwerty keyboard were labeled "type 1", "type 2", "type 3" and "other" respectively. After the response the next cell was displayed.

After the 4th response the next learning block began, followed by its verification phase, followed by the last learning block, and its verification phase. Which kind of cell was displayed as the ten learning cells and two verification cells for each learning block and learning verification phase depended on which level of the learning order factor participants were in. The first type of cell the participants saw was always labeled "type 1", the second "type 2" and so on. Participants never

saw the names type A, type B and type AB.

After the learning verification stage of the last learning block participants entered one final verification phase, to check their memory of each category. After pressing SPACE six cells were displayed for responding as before. Two were type A cells, two were type B cells and two were type AB cells displayed in a random order. The response prompt offered participants a choice between "type 1, type 2 or type 3".

On completing the final verification phase participants entered the test phase. They were instructed that they would see 32 cells. For each cell they were told they would see two close-up shots of that cell. They were asked to wait until they had seen both snapshots before making a decision. Each trial began with a 2 s blank. The first snapshot was then displayed for 3 s, followed by 2 s blank, followed by the second snapshot for 3 s, followed by 0.5 s blank followed by the response prompt "type 1, type 2 or type 3". The order of presentation of each snapshot within a pair was random. Responses were not recorded until the response prompt appeared. There were two blocks of 16 trials. There was no interlude between the two blocks. A block consisted of four presentations of each of the snapshot pairs A, B, AB and A-B in a random order. (In Schyns and Rodet's (1997) original experiment there was only one block of snapshots. A second block was added here to increase the sensitivity of the experiment. As this difference occurred at the end of the experiment, it can make no difference to the results in the rest of the experiment.) After the final response, the experiment ended and participants were thanked for their participation.

Results

The mean proportion of trials correct in the learning verification stage was

high in both conditions: 0.96 and 0.91 for the A→B→AB and AB→A→B conditions respectively. The difference between the two conditions was significant, $t(31)=2.19$, $p<0.05$. The learning was easier in the A→B→AB condition. The mean proportion of final verification trials correct was also high, 0.99 and 0.92 for the A→B→AB and AB→A→B conditions respectively. Again the small difference between the conditions was significant, $t(31)=2.60$, $p<0.05$, with the performance in the elements first condition being more accurate.

Table 8 shows the mean classification of each of the transfer stimuli, for each condition. Data averaged across both transfer blocks is presented (although the pattern remains the same if only the first transfer block is considered, as in the original experiment). Performance on snapshot pairs A, B and AB was high in both conditions. It is performance on the A-B snapshot that is of interest. In the A→B→AB condition the A-B snapshot pair was classified as type AB about half the time, but in the AB→A→B condition, this pair was rarely classified as type AB. This difference in the proportion of type AB responses to snapshot A-B is shown to be significant by a planned t-test, $t(31)=2.78$, $p<0.01$. However, in condition AB→A→B, performance on snapshot AB is poorer than in the A→B→AB condition. If participants remember type AB less well in the AB→A→B condition, then maybe this is why they are reluctant to classify the critical snapshot A-B as type AB.

To rule out this possibility, participants who did not perform significantly above chance in classification of the snapshot pairs A, B and AB in transfer were removed, and the analysis rerun. 1 participant was removed from the A→B→AB condition, and 3 were removed from the AB→A→B condition. Table 9 shows the mean classifications of the transfer stimuli for participants performing above chance

Table 8

Mean proportion of responses for the transfer block of Experiment 9.

		Elements first (A→B→AB)			Compound first (AB→A→B)		
		Response			Response		
		A	B	AB	A	B	AB
Stimulus	A	0.90	0.07	0.03	0.91	0.02	0.06
	B	0.07	0.86	0.06	0.02	0.96	0.02
	AB	0.00	0.04	0.96	0.05	0.19	0.77
	A-B	0.13	0.38	0.49	0.23	0.65	0.13

Table 9

Mean proportion of responses for the transfer block of Experiment 9 for those participants performing significantly above chance on types A, B and AB in the transfer phase.

		Elements first (A→B→AB, n=15)			Compound first (AB→A→B, n=13)		
		Response			Response		
		A	B	AB	A	B	AB
Stimulus	A	0.96	0.01	0.00	0.94	0.03	0.03
	B	0.01	0.93	0.06	0.02	0.95	0.03
	AB	0.00	0.04	0.96	0.01	0.07	0.92
	A-B	0.10	0.38	0.53	0.24	0.62	0.14

on the control snapshot pairs. The pattern described above is maintained. Now performance on the AB snapshot is excellent and about equal in both conditions. The difference between the proportion of A-B snapshots classified as type AB between the two conditions remains significant, $t(27)=2.60$, $p<0.05$.

Discussion

This experiment replicates Schyns and Rodet's (1997) Experiment 2. After training of the same categories, in different orders, two groups of participants gave mutually exclusive categorizations of the same stimulus. This different categorization is attributed to the different groups of participants developing different feature sets.

Experiment 10

Experiment 10 differs from Experiment 9 only in that different Martian Cell stimuli were used. In general the Martian Cell stimuli are slightly smaller, and with less Gaussian blurring. (Gaussian blurring produces an effect similar to an out of focus projector.) The snap shots had a background context added, so the elements were presented on a blank, gray cell background, rather than in isolation on a white screen.

Method

The method for Experiment 10 is the same as for Experiment 9, except for the reduction in blurring of the stimuli and the addition of a blank, gray cell background for the snapshot stimuli. The new stimuli are shown in Figures 30 and 31. New participants took part who had not participated in Experiment 9.

Results

The proportion of correct trials in the learning verification stage was high (0.90 and 0.87 for the A→B→AB and AB→A→B conditions respectively), and

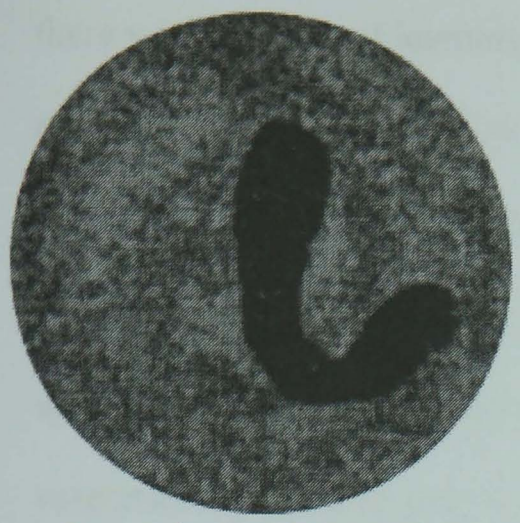
Type A

Type B

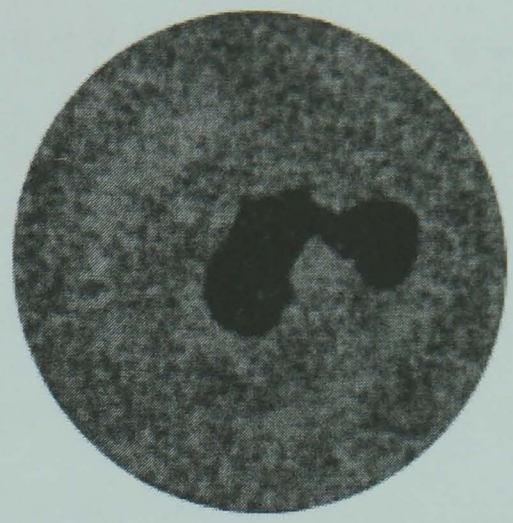
Type AB

Distractor

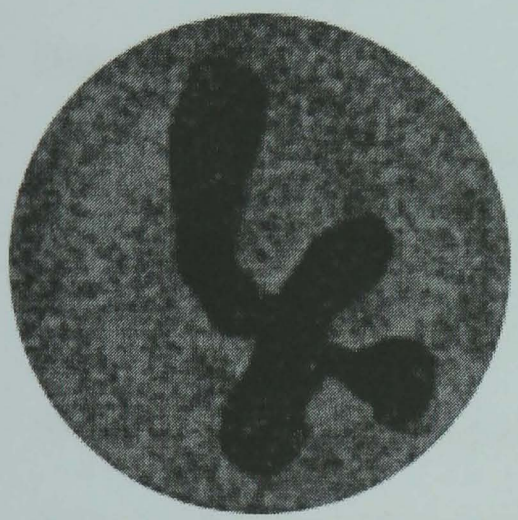
Figure 30. The Martian cell stimuli used in the learning and verification phases of Experiment 10.



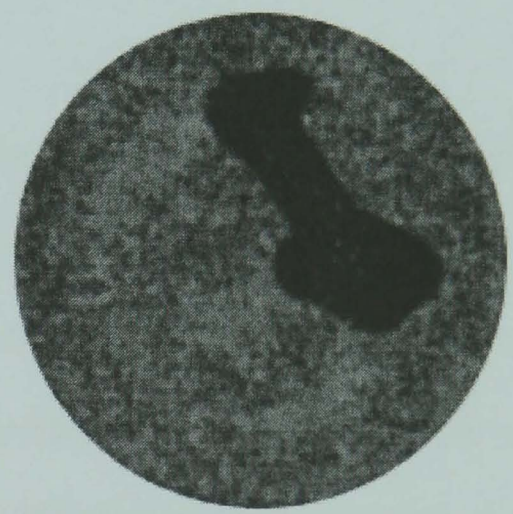
Snapshot of Element a



Snapshot of Element b



Snapshot of Element ab



Snapshot of Distractor Part

Figure 31. The Martian cell snapshot stimuli used in the transfer phase of Experiment 10.

there was no effect of learning order, $t(31)=0.44$, $p>0.05$. Similarly, the proportion of correct responses on final verification trials was high, 0.91, in both conditions.

Results for performance in transfer are shown in Table 10. For the snapshot pairs A, B and AB, performance is high in both conditions. Snapshot pair A-B is almost never classified as type AB in either condition. In fact, out of all the 256 responses to pair A-B, only 4 type AB responses were made. Eliminating participants who did not perform significantly above chance on snapshot pairs A, B and AB (as in Experiment 9) therefore did not alter the pattern of the proportion of type AB responses to snapshot A-B in either condition.

Discussion

The blurring of stimuli and the introduction of a background to the snapshots has completely removed the feature creation effect demonstrated by Schyns and Rodet (1997) and replicated in Experiment 9. Participants who learn in the order A→B→AB no longer categorize the snapshot pair A-B as type AB. The only difference between Schyns and Rodet's (1997) Experiment 2 and this experiment was the stimulus set used. The stimulus sets are, however, very similar, and it is surprising that the feature creation effect is eliminated. The reduction in the blurring of the stimuli should serve to make the relationship between the three types more obvious to participants, as the noise in the stimuli is reduced. The addition of the background context in the snapshots makes it clear that the elements in the snapshots are not joined to other elements. (In the original stimuli, the close-up view could be part of the configural feature.) According to the Schyns and Rodet account, this should not make a difference, as a participant's representation of element ab in the A→B→AB condition is just feature a in the same cell as feature b (as is made clear from their flexible feature encoding simulation, Schyns & Rodet, 1997, p. 691). In

Table 10

Mean proportion of responses for the transfer block of Experiment 10.

		Elements first (A→B→AB)			Compound first (AB→A→B)		
		Response			Response		
		A	B	AB	A	B	AB
Stimulus	A	0.89	0.06	0.05	0.91	0.08	0.02
	B	0.08	0.88	0.04	0.13	0.81	0.06
	AB	0.02	0.02	0.96	0.00	0.08	0.92
	A-B	0.31	0.68	0.02	0.45	0.53	0.02

both experiments participants were told they were seeing close-ups and not isolated parts. Thus in Experiment 9 participants may have believed the A-B snapshots to be separate views of elements a and b joined together. Thus the absence of a feature creation effect here suggests that participants' representations of element ab after A→B→AB training also involves beliefs about the relative location of the two elements.

In fact a stronger claim may be made. Suppose the A→B→AB group and the AB→A→B group have the same fixed feature set {a, b, ab}. Assume that the groups differ instead on their knowledge that element ab is made by joining element a and element b, with the A→B→AB group being more aware that the AB→A→B group. This is a plausible assumption, given that the A→B→AB group can notice previously learned elements being contained in latter elements, but that the AB→A→B group cannot. In Schyns and Rodet's (1997) original Experiment 2, and in Experiment 9, the A→B→AB group can believe snapshot A-B to be two different views of element ab, and classify the snapshot as type AB accordingly. However, the AB→A→B group, being unaware that elements a and b together make element ab, cannot make the same classification. In this experiment the introduction of the background context prevents the A→B→AB group believing that the elements in the snapshot A-B are joined to one another, and therefore they cannot classify the snapshot as type AB. They are prevented from believing that the two elements in the snapshots are joined, because the background context completely surrounds each element, showing the element must be an isolated part of the cell. Thus the feature creation effect observed in Schyns and Rodet's (1997) original Experiment 2, and in Experiment 9, and the lack of the effect in this experiment may be explained using a fixed feature set account.

Partial evidence that this interpretation may be correct is provided by Schyns and Murphy (1994). They trained participants on Martian boulders that contain both element a and element b. In the joined group, these elements were always adjacent. In the separated group, element a and element b were not adjacent. In the joined group, element a and element b were delineated as a single part, but in the separate group they were delineated separately. Participants were then asked to categorize new boulders as either members of the training category or not, and rate them for typicality. Performance on boulders with element a and element b adjacent was the same in both groups, with the examples being categorized as, and rated typical of, the training category. When both elements a and b occurred in the same boulder, but were not adjacent, the joined group was less likely to classify them as and rate them as typical of their training items, even though, to some extent, they could recognize the parts when they occurred alone in a boulder. It seems that although the joined group could recognize the elements a and b alone they required the elements to be joined to make a successful classification.

That the results taken as evidence for feature creation may be explained by an alternative explanation casts potential doubt on the feature creation hypothesis. Experiment 11 was designed to assess whether a feature creation effect may occur with a new type of stimuli, and also to investigate whether participants' knowledge that the compound element is made from a conjunction of the other two elements is predictive of the feature creation effect.

Experiment 11

In Experiment 11 participants learned the features of the three types of line drawings. The elements in these drawings could appear in one of two ways. Half the time the element appeared normally, and half the time the element appeared as a

mirror image. As in previous experiments, one of the elements was the conjunction of the other two elements. The order of learning varied between participants, with half learning the conjunction element, followed by the two part elements, and the other half learning the parts first followed by the conjunction.

In transfer participants saw a stimulus made from a conjunction of one normal part and one mirror image part. This stimulus contains all of the features of the two elements, but none of the features unique to the compound. (This is true if it is assumed that participants learn one feature for both normal and mirror image elements. If alternately, participants learn two separate features for each element, one for normal and one for mirror image presentations, it is still true that the normal-mirror conjunction stimulus does not contain any features unique to the original compound stimulus.) Thus, according to a feature creation hypothesis, participants who learn the elements first represent the compound with two features, one for each element, should classify this new stimulus in the same category as the compound stimulus. However participants who learn the compound first, who may form some features unique to the compound, should be less likely to categorize this new stimulus in the same class as the compound, as the new stimulus does not contain these unique features.

The last part of Experiment 11 is designed to measure participants' awareness that the compound contains the two elements. Of interest is the possibility that the awareness measure might predict participants' classification of the normal element-mirror element conjunction stimulus. The task used to measure awareness was designed to be highly similar to the categorization task where participants were hypothesized to be utilizing this awareness (cf. the information criterion, Shanks & St. John, 1994).

Method

Participants. 32 University of Warwick undergraduates participated for payment.

Stimuli and Design. Two elements were used to construct the stimuli in this experiment, element a and element b (Figure 32A). The elements either appeared normally, or in mirror image. The notation a' is used to represent the mirror image of a. The elements were used to construct three types of training stimuli, A, B and AB (Figure 32B). Stimuli were black lines drawn inside a black outline square. Type A training stimuli contained element a inside themselves. Half contained the element in mirror image. Additional random lines were drawn, by selecting x and y co-ordinates for each end of the line at random, with the constraint that each end of the line was within the outline square. Similarly for type B. Type AB stimuli contained both elements a and b. Sometimes both elements appeared as normal, and sometimes both appeared as in mirror image (denoted using '). Participants never saw one mirror image and one normal element in the same type AB stimulus. Thus the configuration of the elements, either a and b or a' and b', always had the same configural features.

Three kinds of transfer stimulus sets were constructed, a single set, a paired set, and a sequential pairs set (Table 11 and Figure 33). The single set stimuli included the training stimuli. In addition, a new stimulus was included made from elements a and b, but this time so that one element was normal and the other was mirror image, i.e., either a'b or ab'. If participants represented type AB with features a and b, then participants should classify this new stimulus as AB. However, if participants were using a configural feature to represent type AB, then this feature does not appear in this new type, and so participants should classify it as either type A or type B.

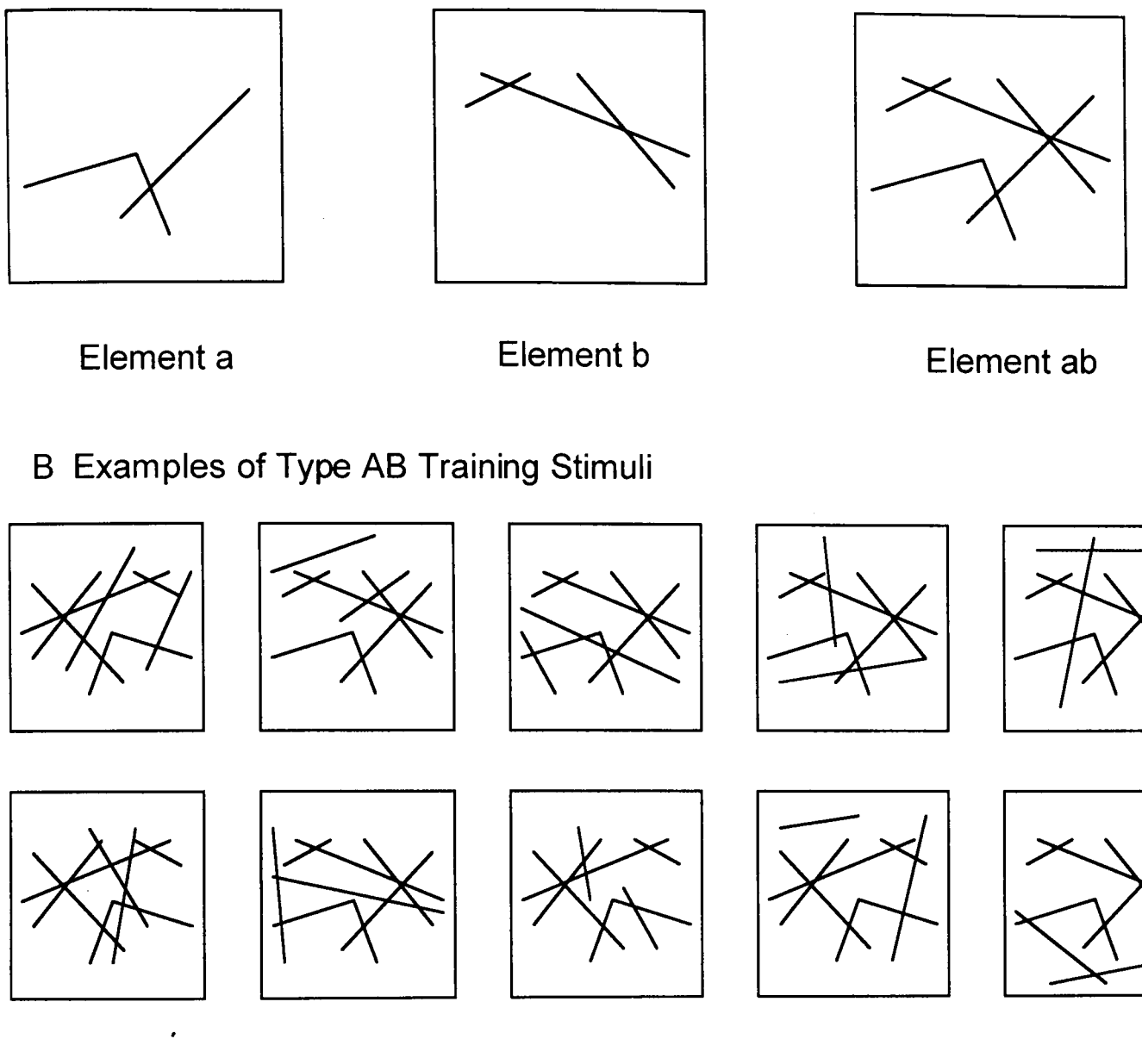


Figure 32. The lines stimuli used in the training phase of Experiment 11. (a) The two elements. (b) Examples of training stimuli.

Table 11

Transfer stimuli from Experiment 11. (a represents element a, b represents element b, a` represents the mirror image of element a, b' represented the mirror image of element b. n shows that addition random lines, or noise, was added to the stimulus. - is used to represent two stimuli being displayed simultaneously, → is used to represent two stimuli, the first displayed before the second.)

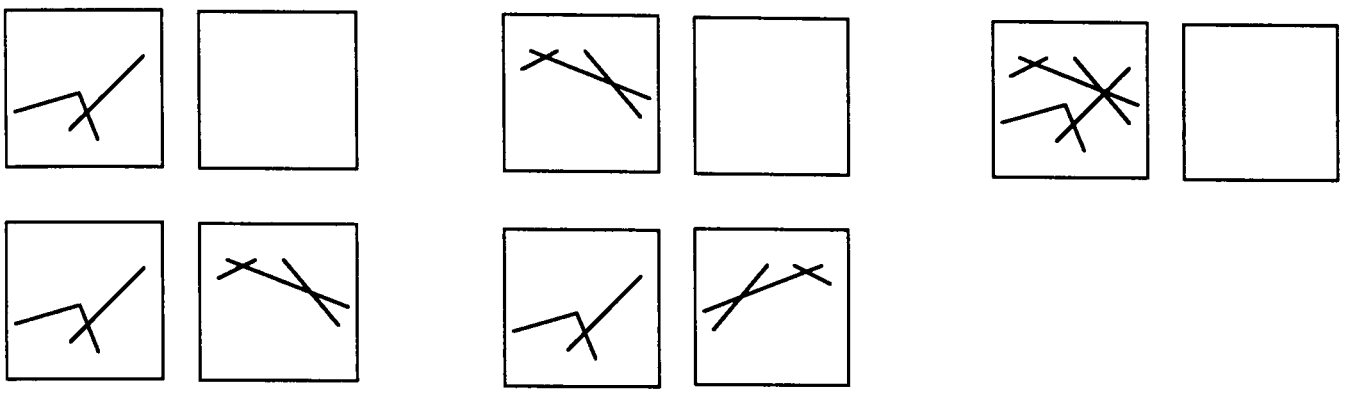
Single set (5 of each)	Pairs set (5 of each)	Sequential pairs set (3 of each)
an	a-n	a'→a
a'n	a'-n	a→a'
bn	b-n	b→b'
b'n	b'-n	b'→b
abn	ab-n	ab→a'b'
a'b'n	a'b'-n	a'b'→ab
ab'n	a-b	a→b'
a'bn	a'-b'	a'→b
	a'-b	b→a'
	a-b'	b'→a
		n→a
		n→a'
		n→b
		n→b`

(table continues)

Single set (5 of each)	Pairs set (5 of each)	Sequential pairs set (3 of each)
		$n \rightarrow ab$
		$n \rightarrow a'b'$
		$ab \rightarrow a'$
		$a'b' \rightarrow a$
		$ab \rightarrow b'$
		$a'b' \rightarrow b$

A Single Example Transfer Stimuli

B Paired Transfer Stimuli



C Sequential Paired Transfer Stimuli

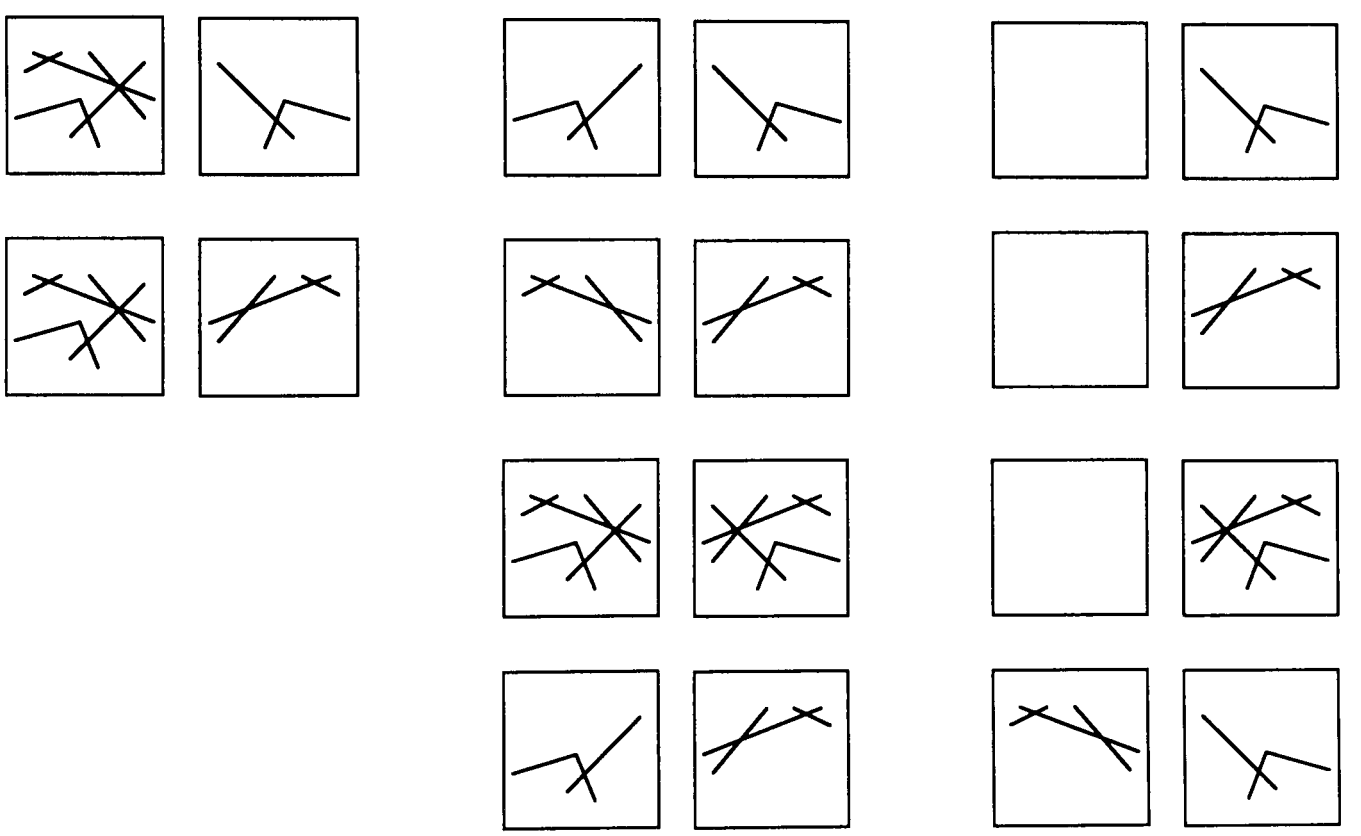


Figure 33. The lines stimuli used in the transfer phases of Experiment 11. Additional random lines have been omitted for clarity, but were seen by participants.

In the paired set, two stimuli were presented on each trial. For each pair stimuli were randomly assigned to a location on the screen on each trial. Five pairs were used. Participants were asked to imagine both stimuli laid one over the other, and to classify the resulting imagined stimulus. The first three pairs were the training examples. The fourth pair, contained element a in one stimulus, and element b in the other, such that when the two stimuli were laid on top of one another the two elements overlapped and formed the conjunction participants had seen in training. The fifth pair was like the fourth, except one element was normal and one was a mirror image. When the two stimuli from this pair were laid one on top of the other, the two elements would not form the conjunction that participants were familiar with. It was hypothesized that participants who learned the parts first, who represented the conjunction stimulus using feature a and feature b, would classify this fifth stimulus pair as type AB. However, participants who learned the conjunction stimulus first, who represent the conjunction stimulus using one feature, feature ab, would not classify this fifth part as type AB, as it does not contain the configural feature ab.

The sequential pairs set was included to measure participants' awareness that one type was made from the parts of the other two. Participants saw the first stimulus, followed by the second. They had to make a binary response indicating whether the second stimulus appeared in mirror image in the first. Of interest are responses to pairs where one of the parts followed the compound stimulus.

Procedure. The first part of the experiment was a pen and paper exercise. Participants were given 3 sheets, each of 10 examples of either types A, B or AB, one after the other. The order of presentation depended on which of two conditions the participants were assigned to. They either learned the compound stimulus first, or

learned the parts first. The order of learning about types A and B was counterbalanced across participants. Participants were instructed that each stimulus contained an invariant part that was either normal or mirror image. They were told to examine the examples and try to spot this part in each example. When they had done this, and shown the part to the experimenter, they were asked to outline the part in each of the 10 stimuli, one at a time, to give themselves a chance to become familiar with the part. Participants were instructed to try to remember the part, as they would be tested on it later.

After all three sheets were complete the transfer phase of the experiment began. In this phase stimuli were displayed on a computer. All participants did the transfer tasks in the order of single stimuli, pairs of stimuli, and sequential pairs of stimuli. On a trial in the single stimulus task a stimulus appeared on the screen until participants responded with one of the category labels from the training stage by pressing one of three keys. (Using small paper stickers the keys q, w, and e on a normal qwerty keyboard were labeled "type 1", "type 2", and "type 3" respectively. For each participant types were named in the order the participant experienced the stimuli.) Stimuli appeared in the center of the screen, the same size they were on the A4 paper in training. After a response there was a blank screen of 500 ms. Trials of each set of stimuli were in a random order. After all the single stimuli were presented, participants were presented with the pairs of stimuli. A trial of the pairs of stimuli task was the same, except two stimuli appeared, one slightly to the left of center, and the other slightly to the right, such that there was a gap the size of one stimulus between the stimuli. After the pairs task, participants moved on to the sequential pairs task. In the sequential pairs of stimuli task, stimuli appeared one after the other, in the center of the screen. The first stimulus was displayed for 1000

ms, with a blank screen for 500 ms before the second stimulus appeared until participants responded. Participants responded by pressing either o or p, labeled "yes" and "no" respectively with small paper stickers.

Results

Two kinds of analyses will be considered. First, the results for the single stimuli and pairs of stimuli transfer sets are presented by condition (A→B→AB and AB→A→B). Planned comparisons were run. Second, the performance in the sequential pairs of stimuli condition will be used to provide a score for each participant measuring how aware they are that the compound stimulus is made up from the two parts. Regression analyses will be used to see if the awareness is a reliable predictor of classification of the ambiguous A'B and A-B stimuli.

Performance on the a, b and ab stimuli in the single stimuli transfer stage was worse in the AB→A→B condition, as observed in Experiments 3 and 4. Many participants could not classify ab stimuli as type AB, after having learning types A and B. Table 12 presents the responses for each stimulus for participants who performed significantly above chance on A, B and AB stimuli. The performance of the A'B stimuli, that contains one mirror image element and one normal element, that is of interest. This stimulus does not contain the configural features of AB, so the AB→A→B group with a single feature ab for stimulus AB should not classify A'B as type AB. Their proportion of AB responses to A'B should be lower than group A→B→AB. The proportion of AB responses to A'B is almost identical for each group, $t(15)=0.03$, $p>0.05$.

The pairs of stimuli transfer set was designed to be like the Schyns and Rodet's (1997) two snapshot procedure used in Experiments 1 and 2. The results for the pairs of stimuli transfer stage are shown in Table 13. Performance on the old

Table 12

Mean proportion of responses for the single stimuli transfer stage of Experiment 11 for participants significantly above chance on types A, B and AB in transfer.

		Elements first (A→B→AB, n=12)			Compound first (AB→A→B, n=5)		
		Response			Response		
		A	B	AB	A	B	AB
Stimulus	A	0.92	0.08	0.01	0.96	0.04	0.00
	B	0.03	0.96	0.01	0.04	0.90	0.06
	AB	0.07	0.04	0.89	0.14	0.04	0.82
	A'B	0.49	0.18	0.33	0.44	0.24	0.32

Table 13

Mean proportion of responses for the pairs transfer set from Experiment 11 for participants significantly above chance on types A, B and AB in transfer.

		Elements first (A→B→AB, n=12)			Compound first (AB→A→B, n=5)		
		Response			Response		
		A	B	AB	A	B	AB
Stimulus	A-N	0.98	0.02	0.01	0.98	0.02	0.00
	B-N	0.02	0.98	0.01	0.00	0.94	0.06
	AB-N	0.05	0.08	0.88	0.08	0.02	0.90
	A-B	0.07	0.01	0.93	0.22	0.08	0.70
	A'-B	0.40	0.11	0.49	0.36	0.16	0.48

training stimuli, A-N, B-N, and AB-N was poorer in the AB→A→B condition. This is consistent with performance on the single stimuli transfer set. Participants failing to perform significantly above chance were eliminated from each condition. The remaining participants in both groups classified the stimulus A-B as type AB most of the time. The proportion of AB responses for stimulus A-B was just significantly higher in the A→B→AB condition, $t(15)=2.145$, $p<0.05$. This is consistent with a feature creation hypothesis. However, the significant difference is due to one participant in the A→B→AB condition almost never classifying stimulus A-B as type AB. The other participants in the A→B→AB condition show a performance similar to those in the AB→A→B condition – the mean proportion of type AB responses to stimulus A-B was 0.83 in the A→B→AB condition with the single participant eliminated, and 0.93 in the AB→A→B condition. As the difference is only marginally significant, and due to just one participant, no more will be made of it. Collapsed across training condition, the A'-B stimulus was classified as type AB less often than stimulus A-B, $t(16)=3.88$, $p<0.01$. This pattern of results is true for both training conditions. The only difference between the A-B and A'-B stimuli is the absence of the configural features in the imagined combination. Therefore that performance differs between these two stimuli suggests that participants in both conditions were aware of the configural feature in stimulus type AB. There is no almost no difference between the proportion of type AB responses to stimulus A'-B between the two training conditions, $t(15)=0.05$, $p>0.05$.

In summary the different orders of learning, either A→B→AB or AB→A→B, has no effect on participants classification of the critical stimuli A'B, A-B and A'-B. However, individual participants do differ on the proportion of AB responses to each of these stimuli. The next analysis investigates whether this

individual variation in classification of these ambiguous stimuli can be explained by participants' awareness that the compound contains the parts.

Table 14 shows the proportion of responses indicating that the 2nd stimulus in each sequence pair appeared as a mirror image in the 1st stimulus. For the control pairs AN→A', BN→B' and ABN→A'B', performance was accurate and approximately equal in each learning condition. For the control pairs N→A, N→B, N→AB, AN→B' and BN→A', performance was also accurate and approximately equal in both learning conditions. The pairs of interest, ABN→A' and ABN→B', designed to measure participants awareness that the compound stimulus contained the parts, performance was intermediate between the two sets of control stimuli. Averaging the results across both critical stimuli for each participant, there is no significant difference in the proportion of yes answers between the two experimental conditions, $\underline{t}(15)=1.18$, $\underline{p}>0.05$ – the different orders of training have no significant effect of the awareness that the compound element in made from the other two elements.

For each participant who performed above chance on the A, B and AB stimuli, and the A-N, B-N and AB-N stimuli, an awareness score was constructed, by taking the average proportion of "yes" responses for sequence stimuli AB→A' and AB→B'. Scatter plots (Figures 34 through 36) show that awareness was a poor predictor of classification of the critical stimuli. Less than 10% of the variability in the proportion of type AB responses to stimulus A'B ($\underline{r}^2=0.090$), A-B ($\underline{r}^2=0.001$), and A'-B ($\underline{r}^2=0.003$) was predicted by this awareness score. Although the correlation coefficient was significantly different from zero, $\underline{t}(15)=4.71$, $\underline{p}<0.0005$, for the A`B stimulus, neither correlation for the A-B and A'-B stimuli was significant, $\underline{t}(15)=0.54$, $\underline{p}>0.05$, and $\underline{t}(15)=0.76$, $\underline{p}>0.05$, respectively. Although there is evidence

Table 14

Proportion of "yes" response in the sequential pairs transfer stage indicating the latter stimuli is contained as a mirror image in the former for participants significantly above change on types A, B and AB in the single and pairs transfer stages in Experiment 11.

	Condition	
Stimulus pair	A→B→AB	AB→A→B
AN→A'	0.81	0.70
BN→B'	0.71	0.80
ABN→A'B'	0.92	0.90
ABN→A'	0.74	0.80
ABN→B'	0.76	0.53
N→A	0.13	0.07
N→B	0.22	0.13
N→AB	0.14	0.30
AN→B'	0.26	0.13
BN→A'	0.13	0.07

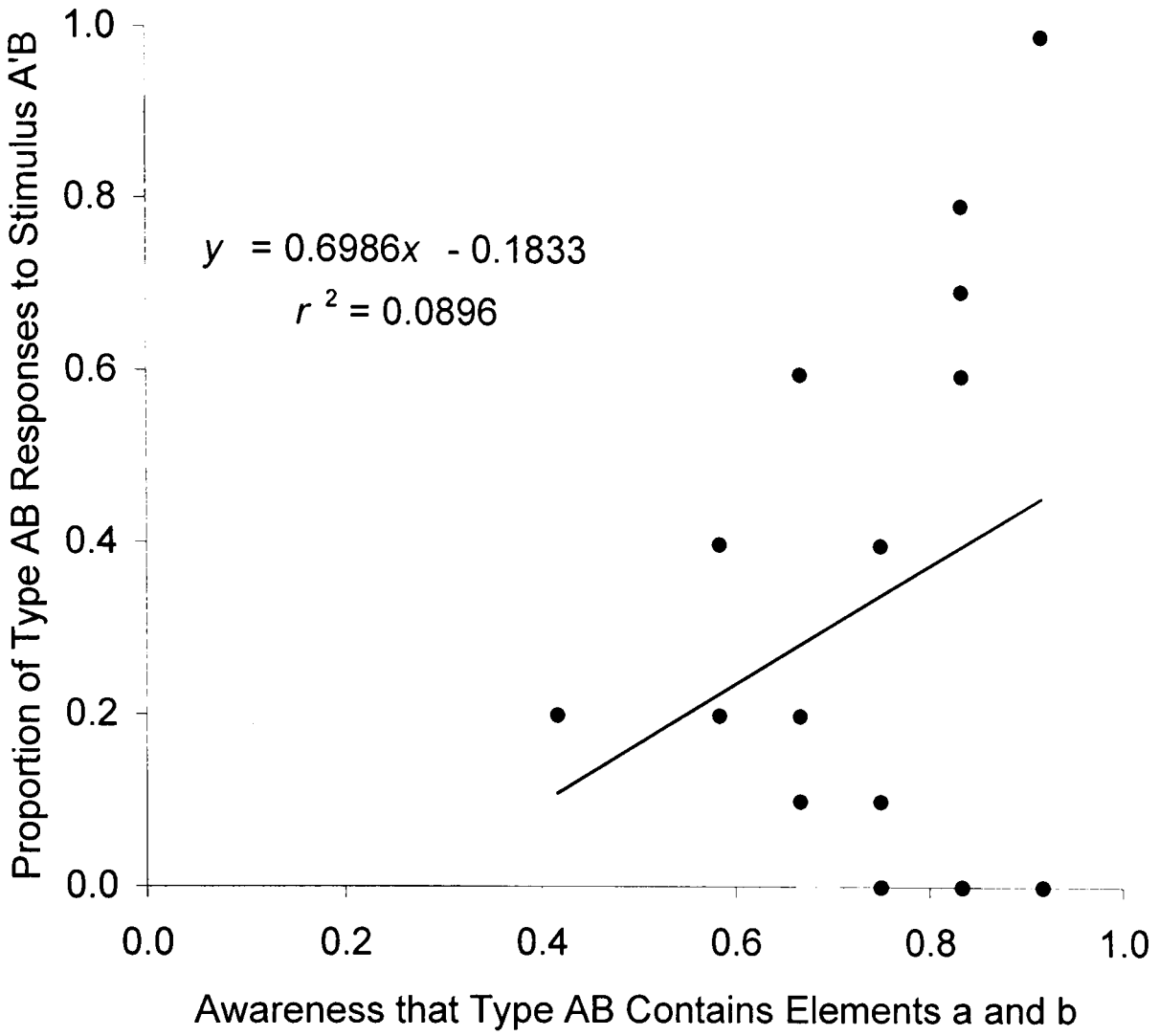


Figure 34. The proportion of type AB responses for the single stimulus A'B transfer stimuli as a function of the awareness that the compound is made from the parts for Experiment 11.

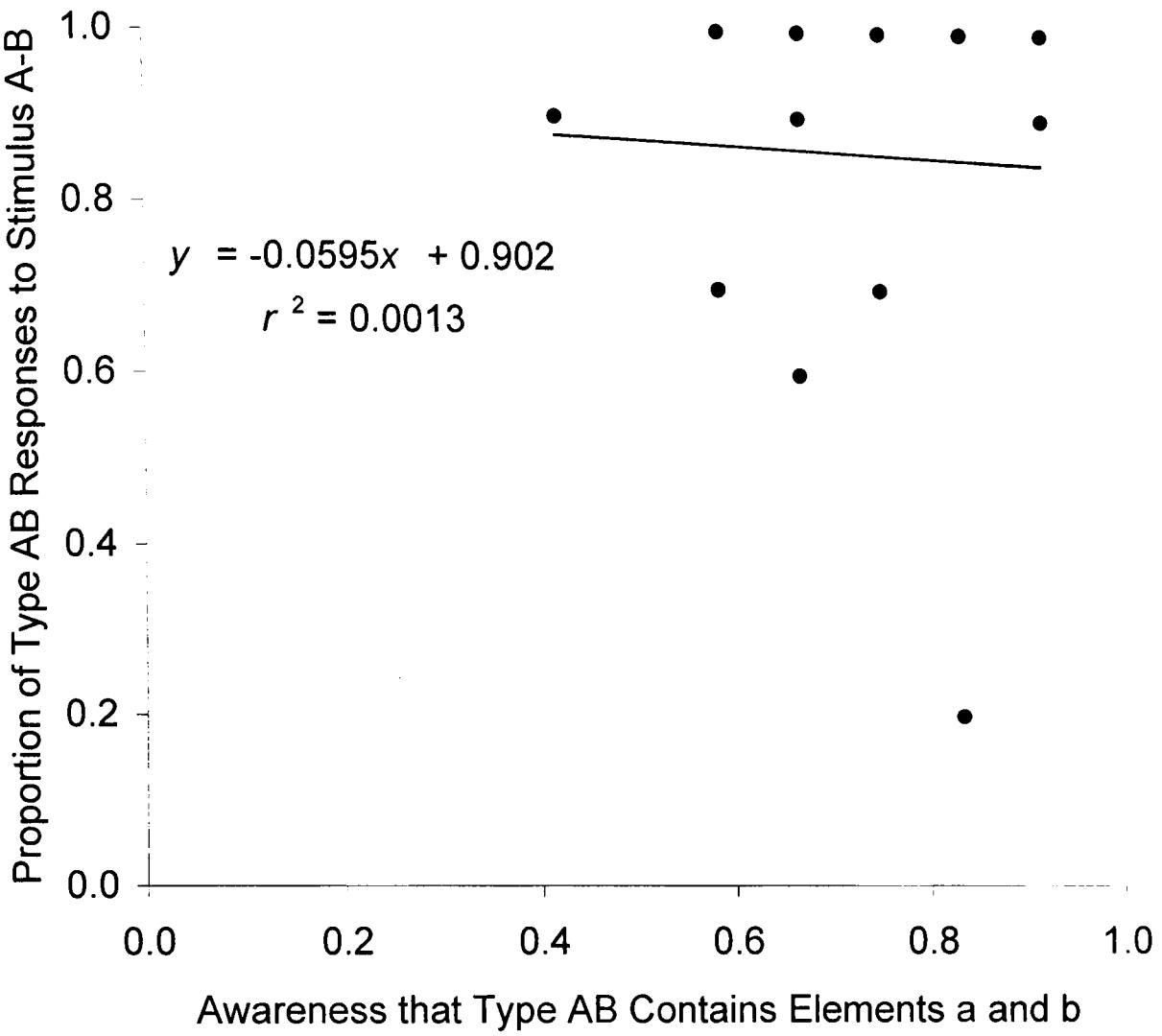


Figure 35. The proportion of type AB responses for the single stimulus A-B transfer stimuli as a function of the awareness that the compound is made from the parts for Experiment 11.

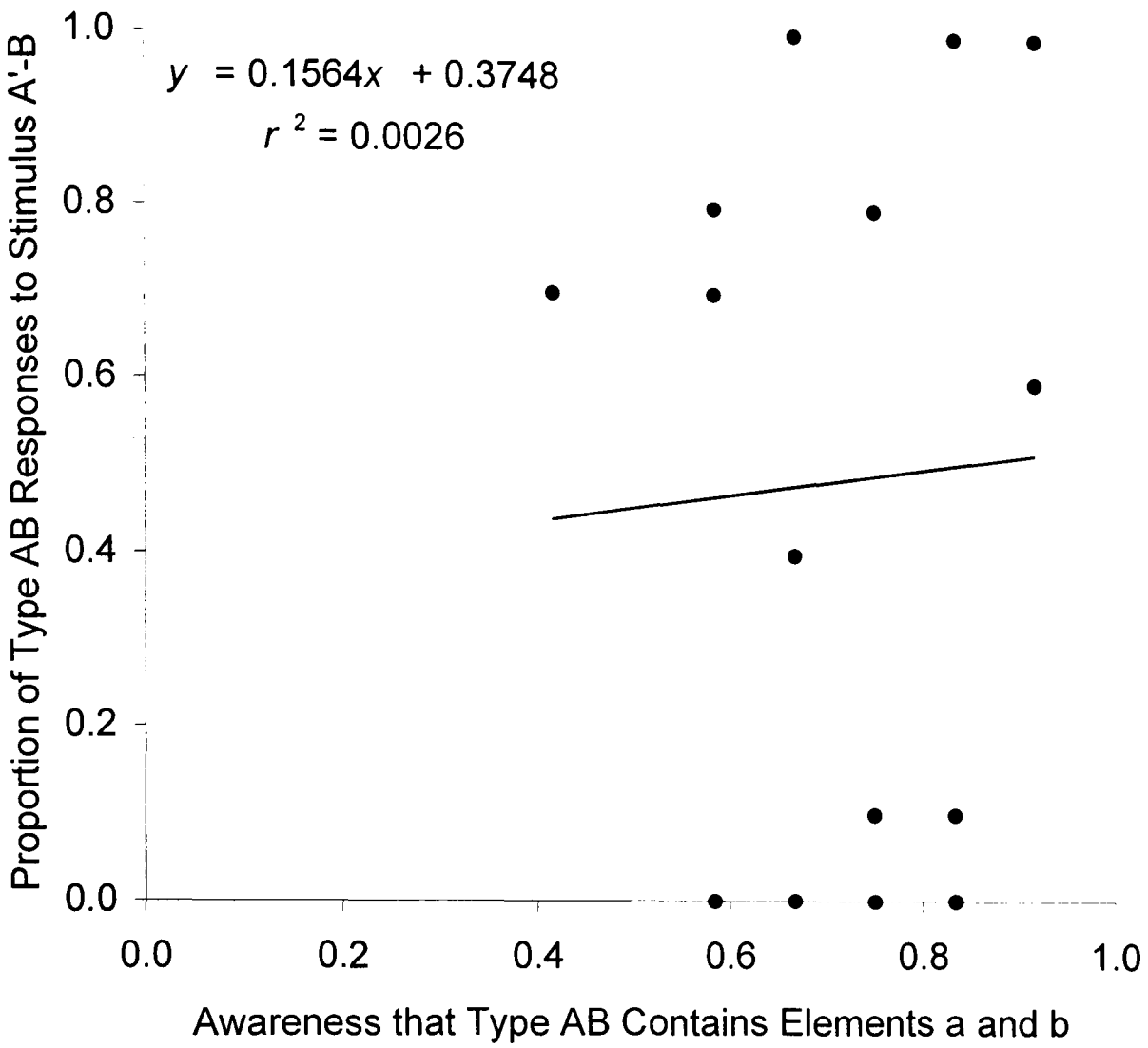


Figure 36. The proportion of type AB responses for the single stimulus A'-B transfer stimuli as a function of the awareness that the compound is made from the parts for Experiment 11.

that the classification of the critical stimuli is influenced by a participant's knowledge that the compound is made up from the parts, the small size of the effect, and absence of a significant effect for other critical stimuli, shows that the large variability in the classification of the critical stimuli is not explained by the experimental manipulations.

Discussion

Participants' classification of the ambiguous stimuli varied greatly. This variation was not predicted by the order of learning the categories. Participants' awareness of the configuration being made up from the parts only predicted a small proportion of variability in categorization of the critical stimulus. Experiment 11 therefore provides no evidence of a feature creation effect, and only slight evidence for the role of awareness. The following experiments aim to demonstrate a feature creation effect, and further to demonstrate that features qualitatively alter the perception of stimuli.

Experiment 12

Experiment 11 failed to demonstrate a feature creation effect. The remaining experiments in this chapter are concerned with investigating feature creation with a new type of stimuli. Experiment 12 uses similar logic to Schyns and Rodet's (1997) original experiment. The Martian cells stimuli are replaced with checkerboard stimuli. Checkerboard stimuli were chosen because they are complex, like Martian cells, and because they are used by other researchers in categorization and perceptual learning (e.g., McLaren, 1997; Wills & McLaren, 1998). The checkerboards contain invariant, category diagnostic patches of black and white squares: elements a and b. Participants learn about three types, type A containing element a, type B containing element b and type AB containing element a above and adjacent to element b.

Participants either learn the single element stimuli first or the compound stimulus first. If participants learn the compound first they may learn features to represent patches of squares across the conjunction of element a and element b. Participants who learn the elements first should have learning of features of the boundary of elements a and b in the compound blocked by their prior learning of features for each element. To allow presentation of the hypothesized element features without presenting features unique to the compound of the elements, the invariant patches in the compound checkerboard are rearranged within a checkerboard (instead of a two snapshot procedure in transfer). This new stimulus, type BA, contains element b above and adjacent to element a. Thus type BA contains the features of each element, but not features unique to the conjunction of element a above element b. Therefore, type BA should be classified as type AB by the group learning the elements first, who do not have any features unique to the conjunction. But the group who learned the compound first, who may have formed features unique to the conjunction, should be less likely to classify type BA as type AB, a type BA does not contain these conjunction unique features.

Method

Participants. 33 students from the University of Warwick took part in the experiment for payment, and did not participate in any other experiment in this chapter.

Stimuli. Two square (4×4) prototype checkerboards were created for each participant. The first prototype board, prototype A, was made from an equal number of black and white squares in random positions. The second prototype, prototype B, was made by randomly selecting half of the white squares of prototype A and changing them to black, and selecting half of the black squares of prototype A and

changing them to white. (This procedure makes the prototypes maximally dissimilar.) A pair of example prototypes is shown in Figure 37.

The prototype boards were used to build four kinds of rectangular (4×8) checkerboard stimuli (Figure 38). The stimuli measured 11 mm across and 22 mm high. Stimulus type A consists of prototype A and random noise checkerboard. For each presentation of a type A stimulus the random noise checkerboard was newly generated. Half the time prototype A made the top half of the rectangular checkerboard, half the time prototype A was presented in the bottom half of the checkerboard. Stimulus type B was as stimulus type A, except prototype A was replaced with prototype B. Stimulus AB was always prototype A above prototype B. Stimulus BA was always prototype B above prototype A. All stimuli were presented on a gray background.

The size of the checkerboards was decided upon so that the chance of learning features that cross the boundary of two adjacent elements should be approximately equal to the chance of learning features within an element. If the boards are two small, then single elements are too small to learn, as the probability of similar elements occurring by chance is too high. If the boards are too large, then the borderline area across two elements is small compared to the area of an element, which means the probability of learning configural features it too low. Pilot work revealed that adding noise in the form of swapping the color of some of the squares made learning checkerboards of this size very difficult for many participants. For this reason, no noise was added.

Design. Participants first learned to discriminate checkerboard types A, B and AB from random noise checkerboards by classifying checkerboards as "type 1", "type 2", "type 3" or "other". A between participants factor varied the order of

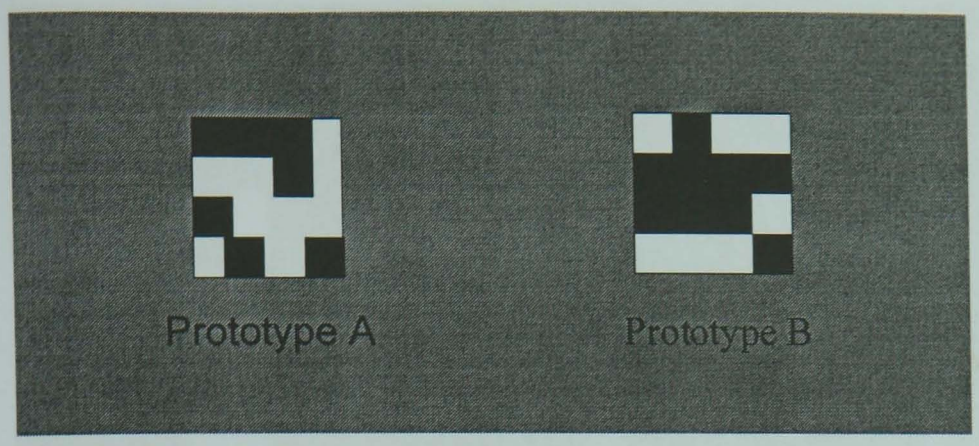


Figure 37. A pair of checkerboard prototypes from Experiment 12.

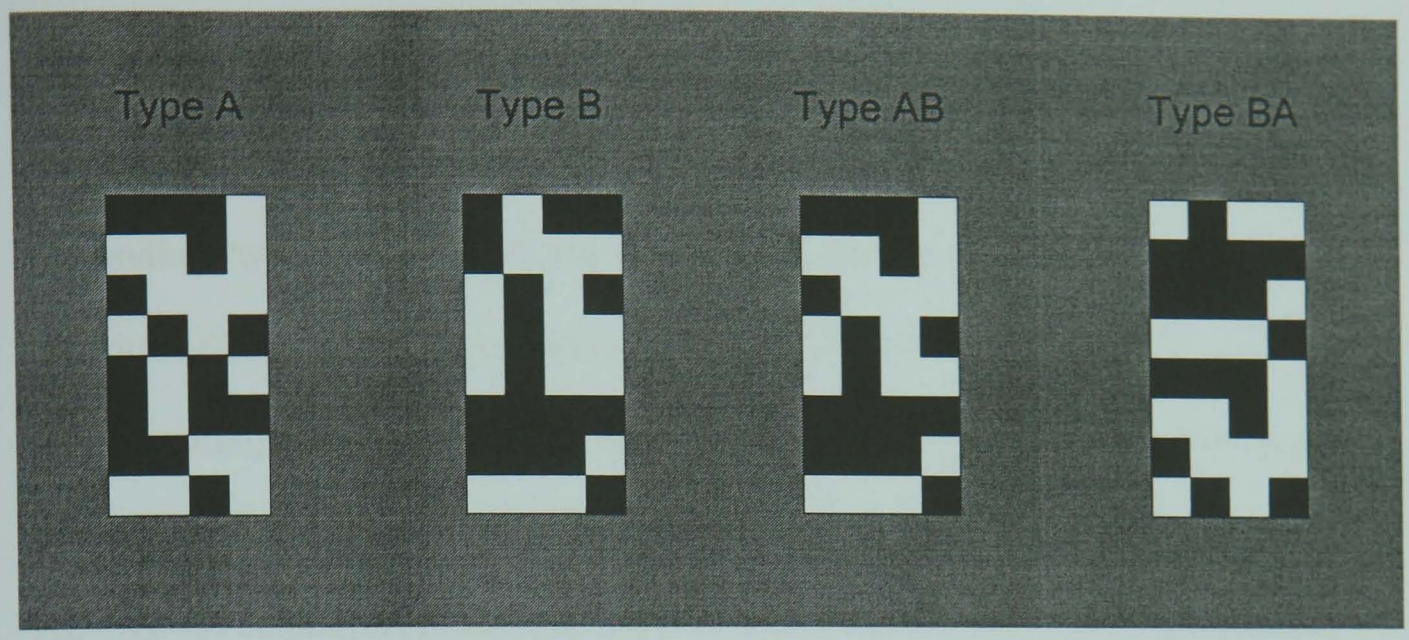


Figure 38. A stimulus set from Experiment 12 (generated from the prototypes in Figure 37).

discrimination learning (Table 15). In the A→B→AB condition participants first learned to discriminate type A checkerboards from random noise checkerboards by responding "type 1" or "other" after each checkerboard. After reaching a set learning criterion of 20 correct categorizations in a row participants moved on to discriminating type B checkerboards from random noise ("type 2" or "other"), and finally type AB checkerboards from random noise ("type 3" or "other"). In the AB→A→B condition participants learned to discriminate type AB checkerboards from noise ("type 1" or "other"), followed by type A from noise ("type 2" or "other"), followed by type B from noise ("type 3" or "other").

They were then asked to classify checkerboard types A, B, AB and BA and random noise checkerboards as either "type 1", "type 2", "type 3" or "other". It was predicted that participants in the AB→A→B condition may learn to classify type AB checkerboards using features on the boundary between prototype A and prototype B. Learning of these features should be blocked in the A→B →AB condition, as participants have already learned features a and b and these should be sufficient for discrimination. The absence of the features representing the boundary between element a and element b in the BA stimulus should reduce the likelihood of participants in the compound first condition classifying type BA as type AB compared to the elements first group.

Procedure. The experiment took place in a quiet room. Participants were seated in front of a computer and the keyboard and monitor were adjusted as necessary. Participants read the instructions from the computer screen and were given an opportunity to question the experimenter. They were told that they could learn to identify which checkerboard belonged to which type by learning patches of squares that would appear somewhere in each type. Three example boards were

Table 15

The design of Experiment 12.

	Training Discriminations			Transfer
Condition	Type 1, other	Type 2, other	Type 3, other	
A→B→AB	A, noise	B, noise	AB, noise	A, B, AB, BA, noise
AB→A→B	AB, noise	A, noise	B, noise	A, B, AB, BA, noise

displayed below the instructions. The example prototype was randomly generated. and was not used during the experiment. Participants were shown where the prototype pattern appeared in all three boards. The experiment then started with instructions for the first learning block. Participants were told they would see checkerboards one after the other, and that they should categorize each one as either "type 1" or other. Although they would have to guess at first, they were told they should eventually be able to learn by paying attention to the correct answer. Each trial in the learning block was proceeded by 500 ms of blank gray screen. A rectangular checkerboard stimulus was then displayed in the center of the screen. After 2000 ms a 15 mm high "?" prompt was displayed underneath the board in until the participant responded pressing keys labeled "type 1" or "other". The 2000 ms pause was to encourage participants to look at all parts of the checkerboard before responding. (Using small paper labels, the keys z, x, c, and v on a normal qwerty keyboard were labeled as "type 1", "type 2", "type 3" and "other" respectively.) The stimulus remained on the screen and the prompt was overwritten with the correct response "type 1" or "other", again 15 mm high. The feedback lasted 2000 ms, after which time the screen was cleared. The next trial then began automatically. On each trial a type 1 checkerboard or a newly generated random noise checkerboard was selected for presentation. The selection was random. After participants had made 20 correct responses in a row they moved onto the next learning block. The second and third learning blocks were the same as the first, except that the type of checkerboard and labels used were changed.

After the third learning block the transfer block began with instructions telling participants that they would see more checkerboards to classify. They were told that the checkerboards could belong to any type, and to look at all the parts of

each board carefully before making a response. The format of each trial was the same as the learning block, except the feedback was omitted. There were 50 trials. 10 of each of checkerboard types A, B, AB, and BA and 10 random noise checkerboards. The trials were in a random order.

Results

33 undergraduates participated. One participant in the AB→A→B first condition failed to complete the training phase after one hour, and was excluded from the following analysis. This leaves 16 participants in each of the two conditions of the experiment. The mean number of trials to criterion is displayed in Table 16. Both conditions show an effect of practice, with less trials to criterion being taken when learning later types. The individual variation in blocks to criterion is large.

The performance in the transfer phase of the experiment is displayed in Table 17. In the A→B→AB condition, performance on the old training items type A, B and AB is good, at about 80% accurate. The new item type BA is classified as types A, B and AB about equally. In the AB→A→B condition performance on type A and B is good, but performance on the compound stimulus, type AB, is poor. Type AB is classified as type A or type B most of the time. Because of this, the lower rate of classification of type BA as type AB in the AB→A→B condition compared to the A→B→AB, $t(31)=2.11$, $p<0.05$, cannot be taken as evidence for a feature creation effect. It seems more parsimonious to account for this result by noting that if participants do not classify the type AB stimulus as type AB, because they cannot remember it, then they will not be able to classify the type BA stimulus as type AB either. Unfortunately eliminating participants who did not perform significantly above chance on the old training items in transfer (i.e.. 6 or more correct responses out of 10), demonstrating memory for the old training items, eliminates all but two

Table 16

Mean number of trials to criteria in each of the training blocks for Experiment 12.

Condition	Type A	Type B	Type AB
A→B→AB	70	56	27
AB→A→B	50	48	41

Table 17

Mean proportion of responses in the transfer stage of Experiment 12.

		Elements first (A→B→AB)				Compound first (AB→A→B)			
		Response				Response			
		A	B	AB	Noise	A	B	AB	Noise
Stimulus	A	0.82	0.01	0.06	0.11	0.74	0.00	0.18	0.08
	B	0.01	0.82	0.06	0.10	0.02	0.83	0.04	0.11
	AB	0.06	0.11	0.80	0.02	0.40	0.28	0.31	0.01
	Noise	0.04	0.06	0.01	0.89	0.04	0.03	0.13	0.81
	BA	0.33	0.33	0.30	0.05	0.24	0.66	0.07	0.03

of the participants from the AB→A→B condition. (12 participants remain in the A→B→AB condition.) Further detailed analysis is therefore not possible.

Discussion

Participants in the A→B→AB condition demonstrated good memory for each of the training categories in test. However participants in the AB→A→B condition were not able to maintain a memory for type AB. Because of this, their lower proportion of AB classifications to the type BA stimulus cannot be taken as evidence for a feature creation effect, but may be explained simply be observing that if participants can't remember type AB they are unlikely to classify anything as type AB. Better memory in the A→B→AB condition is in fact consistent with a feature creation hypothesis. According to the feature creation hypothesis, in the A→B→AB condition only two features need be maintained in memory, {a, b}, but in the AB→A→B condition three features, {a, b, ab}, must be maintained.

Experiment 13

Experiment 13 differs from Experiment 12 only in a slight change to the training stimuli. In this experiment the stimuli were manipulated to reduce the likelihood of learning of a single conjunction feature for the AB stimulus. In other words, the manipulation was introduced to disrupt the formation of features across the boundary of elements a and b in the AB stimulus. This should make the task easier for the AB→A→B group, who would now only have to maintain two features in memory. If the manipulation that prevents the learning of a conjunction feature, performance in transfer should now be the same across both groups A→B→AB and AB→A→B. Further all participants in this experiment should be more likely to categorize the transfer item A-B as type AB, as all participants should be less likely to have learned a single configural feature for stimulus AB, and will instead

represent it using the two elements which are also present in stimulus A-B. The manipulation used was the introduction of colors to the checkerboards in the learning phase. In the learning phase, the top and bottom of the checkerboards were given different colors. One half was colored pink and red instead of white and black respectively, and the other half light blue and dark blue instead of white and black respectively. Thus the patterns of light and dark was the same for stimuli in training and test.

Method

The method is the same as for Experiment 12, except for differences described here.

Participants. 32 students from the University of Warwick took part in the experiment for payment, and did not participate in other experiments in this chapter.

Stimuli. The stimuli were generated in the same way as Experiment 12. The only difference was that colors were added to all the stimuli in the training phases. The top and bottom halves of the checkerboards were given different colors. One half was red and pink, and the other dark blue and light blue instead of black and white respectively. Stimuli in the transfer phase of the experiment were black and white as in Experiment 12. Although the old colored training stimuli appeared in black and white in the transfer phase, participants could still identify them as the pattern of light and dark squares was still the same. On each trial in the training phase the assignment of colors to the top or bottom of the checkerboard was random. The diagnostic part of each type of checkerboard could appear in any color.

Procedure. The procedure was the same as Experiment 12, except additional instructions were given. Participants were told to ignore the colors in the training stage, and just pay attention to the pattern of light and dark squares, as this is what

would help them learn which type was which. Participants were told that the checkerboards in the transfer phase would be black and white versions of the training checkerboards, and not to worry about this, but just to categorize them as best they could.

Results

All 32 participants completed the experiment. The mean number of trials to criterion is displayed in Table 18. Both conditions show an effect of practice, with less trials to criterion being taken when learning later types. Compared to Experiment 12, participants took approximately half the number of trials to criterion. suggesting the color manipulation did indeed make learning easier.

The performance in the transfer phase of the experiment is displayed in Table 19. The pattern of performance on the old transfer items the same as that observed in Experiment 12. In the A→B→AB condition, performance on the old training items a, b and ab is good, at about 80%. In the AB→A→B condition performance on the elements a and b is good, but performance on the compound stimulus, type AB, is poor. Type AB is classified as type A or type B most of the time.

The proportion of type AB response to the new transfer item type BA is high in condition A→B→AB. Participants in this condition classified type BA as type AB most of the time, and is significantly higher than in the AB→A→B condition, $t(31)=4.90$, $p<0.0001$. However, as in Experiment 12, this cannot be taken as evidence for a feature creation effect, because of the poor performance of participants in the AB→A→B condition on the old transfer item type AB. All but three participants in this condition showed the poor performance on this item. Eliminating these participants does not leave enough in the condition for a sensible comparison, but the difference between the two conditions was in the direction

Table 18

Mean number of trials to criteria in each of the training blocks for Experiment 13.

Condition	Type A	Type B	Type AB
A→B→AB	29	26	23
AB→A→B	32	27	25

Table 19

Mean proportion of responses in the transfer stage of Experiment 13.

		Elements first (A→B→AB)				Compound first (AB→A→B)			
		Response				Response			
		A	B	AB	Noise	A	B	AB	Noise
Stimulus	A	0.74	0.00	0.16	0.10	0.75	0.05	0.12	0.08
	B	0.01	0.83	0.01	0.16	0.02	0.81	0.09	0.08
	AB	0.01	0.09	0.89	0.01	0.32	0.32	0.30	0.06
	Noise	0.01	0.00	0.01	0.98	0.03	0.02	0.06	0.89
	BA	0.15	0.09	0.67	0.09	0.38	0.43	0.12	0.07

predicted, with a very low proportion of type AB responses to type BA in condition AB→A→B but not A→B→AB.

One last prediction remains to be tested. It was hypothesized that the introduction of different colors for each part of the stimulus would encourage the type AB stimuli to be learned as two halves, in both conditions. Thus the proportion of type AB responses to type BA stimuli should be higher in this experiment than Experiment 12, as fewer participants should form the configural feature ab. Performance on the new transfer item, type BA, does differ from Experiment 12. As predicted the proportion of type AB responses to type BA stimuli is larger than in Experiment 12, $t(63)=2.67$, $p<0.001$. This comparison is across experiments, and so must be viewed with care, as participants were not randomly assigned to either experiment.

Discussion

The pattern of results is similar to that observed in Experiment 12. Participants in the A→B→AB condition classified type BA stimuli as type AB significantly more often than participants the AB→A→B condition. However, as in Experiment 12, participants in condition AB→A→B did not successfully maintain a representation of the type AB stimuli over learning of the intervening types. Thus although the difference in classification of the type BA stimuli is consistent with a feature creation hypothesis, it is more likely that forgetting is responsible for the difference.

Experiment 13 differed from Experiment 12 only in the introduction of coloring the stimuli in the learning phase, to disrupt the learning of features of the boundary of the conjunction between elements a and b in type AB stimuli. Thus the proportion of type AB response to type BA stimuli should be significantly higher in

this experiment as participants should have been less likely to learn conjunction features. This was indeed the case. Thus although neither experiment alone provides compelling evidence for feature creation, this comparison across the two experiments is suggestive of a feature creation effect.

What is needed is a paradigm to investigate feature creation that does not involve participants having to remember features over long blocks of intervening trials. Experiment 14 provides such a paradigm.

Experiment 14

Experiment 14 has three stages in which participants classify 10 × 10 square black and white checkerboards. Each checkerboard can be considered then as having four quadrants, each 5 × 5 squares. In the first stage, participants learn to classify two types: types A and B. Type A contains an invariant 5 × 5 square area of checkerboard, element a, appearing in one of the quadrants. The remaining quadrants are filled with a random pattern of black and white squares that varied from trial to trial. Likewise, type B checkerboards contains a different 5 × 5 square area of checkerboard, element b. The quadrant the diagnostic feature appears in varies from trial to trial for each type.

When participants reach criterion they move on to learn three new types: type AB, type CD and type E-F. Each type now contains two diagnostic quadrants. Type AB contains the previously learned elements a and b, in random locations, with the constraint that element a always appears above element b. Although this two quadrant feature could be learned as a single unit, it was hypothesized that prior learning of the two parts of the large feature should block learning of a configural ab feature. Type CD contained two new elements, c and d, that appeared at random with the constraint that c always appeared above d. Here, with no prior experience, it was

hypothesized that participants would learn the two joined features as a single configural unit, feature cd. Type E-F also contained two new elements. e and f, appearing in random locations, with the constraint that e never appeared directly above or below f. Because the two elements in type E-F did not consistently appear in the same location relative to one another, participants must learn two separate features to represent type E-F.

In the final stage of the experiment participants classified briefly presented checkerboards as type AB, type CD or type E-F. They saw old examples of type AB, type CD and type E-F, and, in addition, three new types of stimuli. The new types were a checkerboard with element b above element a (type BA), a checkerboard with element d above element c (type DC), and a checkerboard with element f above element e (type FE). If participants do indeed have a configural representation feature cd, then they should not classify DC as type CD, because it does not contain the feature cd. In contrast, participants should classify types BA and FE as types AB and EF respectively.

Method

Participants. 24 University of Warwick undergraduates took part for payment. Participants had not taken part in any other experiment in this chapter.

Stimuli and Design. Seven 5×5 checkerboard prototypes were created for each participant. Each of a prototype's 25 squares was randomly assigned to be either black or white with equal probability. One of the prototypes was used as the example feature in the instructions. The remaining 6 prototypes were used to construct the stimuli. These six prototypes will be referred to as elements a through f.

The checkerboard stimuli displayed to participants always measured 10×10 squares. Either no quadrants or one or two quadrants were replaced with a or b

elements. In the initial stage of the experiment participants learned to categorize three types of stimuli, type A, type B, and other completely random checkerboards. Type A stimuli always included element a in one quadrant. The quadrant was selected at random on each trial. The remaining squares of the checkerboard were randomly assigned to be black or white on each trial. Thus the only invariant part of the stimulus is element a. Likewise for type B. The completely random checkerboards were included to ensure participants learned both types A and B. When participants reached criterion on these boards (over 90% correct) they learned three new checkerboard types, type AB, type CD and type E-F. Type AB boards element a above element b. The two quadrants could either appear in the left part of the board or the right. The remainder of the board was random, as previously described. Type CD boards contained the (previously unseen) elements c and d in the same arrangement, with the remainder being random. Type E-F boards contained elements e and f, in random quadrants, with the constraint that one never appeared above the other. (The dash denotes that the features were separate.) The remaining two quadrants were random. It was hypothesized that a configural feature, cd, would be learned to represent the conjunction of elements c and d. Learning of the analogous feature for the AB boards, ab, would be blocked by prior learning of features for a and b separately in the first part of the experiment. Separate features for elements e and f would have to be learned as the elements occur in many different locations relative to one another.

In the final stage of the experiment participants classified the previously described types AB, CD and E-F. In addition they classified three new types of stimuli, BA, DC, and FE. Checkerboard BA contained element b above element a, with the remaining half of the board being random. Likewise for type DC and FE. As

BA and FE both contain the same features as types AB and E-F respectively, these could be classified. However, type DC does not contain the configural feature cd, and thus should not be classified as type CD.

Procedure. The experiment took place in a quiet room. Participants were seated in front of a computer and the keyboard and monitor were adjusted as necessary. Participants read the instructions from the computer screen and were given an opportunity to question the experimenter. As in the previous checkerboard experiments they were instructed that boards belonging to the same type all had a common invariant patch, and that they should try to learn this patch. Three example boards were displayed below the instructions. The example prototype was randomly generated, and was not used during the experiment. Participants were shown where the prototype pattern appeared in all three boards. The experiment then started with the first learning block. Participants were told that they would see two types of checkerboard mixed in with some random boards. They were told that they would have to guess which type was which at first, but that by paying attention to the feedback after each board they should be able to learn. They were instructed to try to be as accurate as possible. Each trial was proceeded by 500 ms of blank gray screen. A square checkerboard stimulus was then displayed in the center of the screen. The stimulus remained on the screen until the participant responded with keys labeled "type A", "type B", or "other". (Keys z, x, c, v, b and n on a normal qwerty keyboard were labeled "type A", "type B", "type C", "type D", "type E" and "other" respectively.) Below the checkerboard the correct response was displayed as feedback, either "A" or "B" or "other", in 15 mm high text. The feedback lasted 2000 ms, after which time the screen was cleared. The next trial then began automatically. Blocks of 27 stimuli, equal numbers of type A, type B and random

stimuli, were repeated until participants completed a block with no more than two mistakes. The random stimuli were included to ensure participants learned both type A and type B, and did not just learn one type.

After reaching criterion instructions for the second stage of the experiment were displayed. Participants were told that now they would learn three new types. They were told that now each type contained two parts, and they should try to learn both for each type. They were told that one of the new types contained both parts from the first stage. The format for a trial was the same as in the previous stage, except the stimuli and feedback labels were different. Participants were presented with blocks of 20 stimuli. In each block there were 5 type AB checkerboards, 5 type CD checkerboards, 10 type E-F checkerboards, and 5 completely random checkerboards. There were twice as many E-F checkerboards as any other type, as pilot work demonstrated that these were harder to learn than the other types. Blocks were repeated until participants made no more than two errors in a block. When participants reached criterion, the instructions for the third and final stage of the experiment were displayed informing participants that they would see boards appear briefly, and they should try to categorize them as in the second learning stage. A trial began with a fixation cross for 1000 ms, in the center of the screen. A checkerboard was then displayed for 650 ms before being covered over with a mask of random black and white pixels larger than the checkerboard. The purpose of this manipulation was to prevent ceiling effects in the classification of checkerboards. Participants responded with one of the labels "type C", "type D", "type E", or "other". They were instructed to respond other if they were just guessing which checkerboard had been displayed, to reduce chance correct responses to any of the types of checkerboards. 70 trials of checkerboards were run, 10 of each of types AB,

CD, E-F, BA, DC, FE, and 10 random checkerboards in a random order. After the last trial the experiment ended.

Results

Six participants failed to complete the experiment in the time allowed. Their results are excluded from this analysis. On average the remaining participants reached criterion after a mean of 3.7 and 4.7 blocks for the first and second learning phases respectively. Of interest is performance in the transfer stage (Figure 39). For the new items a correct response is defined as responding with the label of the type whose elements it contains. The brief exposure of the checkerboards ensured performance on the old training items was below ceiling. Of interest is the comparison between an old item and the corresponding new item, where the two elements appeared in a novel configuration. In the elements condition, (types E-F and FE), where participants were hypothesized to have a feature for each element e and f, performance is about equal for the two types. In the compound condition (types CD and DC) DC is classified as CD less often than CD is. This is as predicted. If participants have a feature cd that they use to represent type CD, they should not classify DC as CD, as DC does not contain feature cd. In the blocked condition, where participants experienced types A and B before learning about type AB, the pattern of results is intermediate between the other two conditions. This supports the notion that learning types A and B before AB partially blocks the formation of feature ab.

This descriptive pattern of results is confirmed with the following inferential statistics. As a change in proportion is to be compared across 3 conditions, the log odds method is used (Allerup & Elbro, 1998). This method is superior to the more common 3×2 ANOVA (stimulus type $\times$ swapping of elements) that might be run

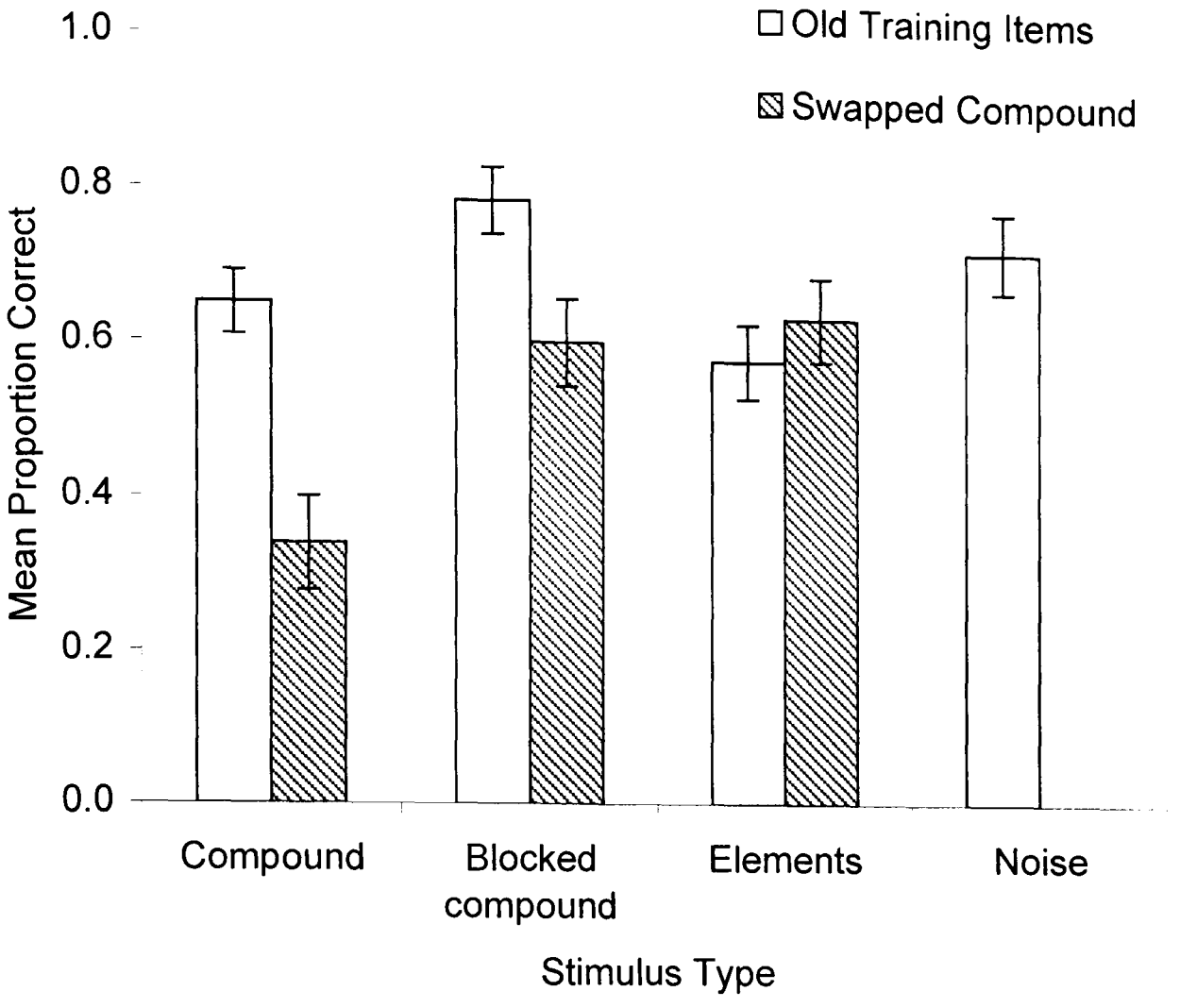


Figure 39. The mean proportion of correct responses for the transfer stimuli in Experiment 14. (Error bars are standard error of the mean.)

here, where an interaction would be taken as evidence of a feature creation effect, because the log odds methods controls for the potential floor and ceiling effects. For each participant three log odds scores were calculated, one for each condition. Odds were calculated for each of the possible six test stimuli (three types × swapping elements over, not including the random checkerboards). The odds-value is defined as the probability of an event happening divided by the probability of it not happening. The probabilities were estimated using the proportion of correct responses for each stimulus type. For each of the three stimulus types, log odds were calculated, defined as the natural logarithm of the quotient of the odds. Each log odds score therefore reflected, for a given stimulus type, the difference between the proportion correct when the stimuli appeared as in training and the proportion correct when the stimulus appeared with the two elements in novel locations, i.e., the difference in height of each pair of adjacent bars in Figure 39. A one way ANOVA revealed a significant difference in this log odds score between the three conditions, $F(2, 34)=22.28$, $p<0.0005$ (Huynh-Feldt $\varepsilon=0.96$). (This would be equivalent to a significant interaction in the 3 × 2 ANOVA.) t-tests showed that the log odds scores for the for the compound and blocked conditions differed significantly, $t(17)=2.41$, $p<0.05$. The difference between the log odds scores for the compound and elements conditions was also significant, $t(17)=7.45$, $p<0.000005$, as was the difference between the blocked and elements conditions, $t(17)=3.70$, $p<0.005$.

Discussion

Performance in transfer was as predicted by the feature creation hypothesis. In type E-F boards elements e and f did not appear in consistent locations relevant to one another. Participants were hypothesized, therefore, to form two features, e and f, to represent each element. Consistent with this interpretation both type E-F and FE

stimuli were classified as type EF in transfer as both types contain features e and f. In type CD stimuli, the elements c and d did appear in a consistent configuration across stimuli, and therefore participants were hypothesized to learn a single feature cd to represent this configuration. Participants classified type DC stimuli as type CD less often than type CD stimuli, consistent with a feature creation hypothesis, as type DC stimuli does not contain feature cd, but type CD does. Intermediate between these two patterns of responding was performance on type AB and type BA stimuli, where although elements a and b did appear in a consistent location relative to one another, prior experience with each element singly was hypothesized to create features for each element, partially blocking the formation of a feature for the conjunction of the elements.

Experiment 15

A second interpretation of the result from Experiment 14 exists which does not involve feature creation. Instead the explanation is based on the change in location of the elements between training and transfer. Experiment 15 investigates this alternate explanation. If participants were impaired at recognizing elements in novel locations, then they might be worse on stimulus DC than either of the other two transfer stimuli, BA and EF. In more detail, in the compound condition of Experiment 14, the test stimulus, DC, had element d at the bottom of the stimuli, and element c at the top of the stimulus. Participants had never seen either of the elements in these locations before. For the other two test stimuli, BA and FE, participants had seen the elements of these stimuli in the location they appear in in the test stimuli before. This confound may therefore explain the reduced performance in the DC condition compared to the BA and FE conditions.

Previous experimental work in visual search suggests that this alternate

explanation may be valid. Treisman, Vieira and Hayes (1992) have demonstrated that a stimulus that appears in more frequently in certain locations is better detected in these locations. Further, and of importance here, is that the effect is specific to the particular stimulus that appears frequently in that location, and not just any stimulus appearing in that location.

A second reason related to consider the effects of element location is predicted by the feature creation hypothesis. Consider an element that consistently occurs at the top of a checkerboard. The top part of the element always occurs next to the gray background. The bottom part of the element may appear next to squares that change color on each trial. This has the effect of changing the appearance of the bottom part of the element from trial to trial. Thus it will be easier for participants to learn a representation of the top part of the element than the bottom part. In transfer, when the element appears in a new location at the bottom of the board, the easy-to-learn top part of the element is now adjacent to black and white squares, and will therefore be harder to recognize. Thus, this feature creation account makes the same predications as the location account.

Experiment 15 is designed to test these alternate explanations. Participants will learn to classify two types of checkerboard, each type containing an element in a consistent location. In transfer, two new checkerboards will be introduced, where the elements appear in non-consistent locations. If there is a significant effect of changing the location of a feature then performance should be worse on these new items.

Method

Participants. 7 University of Warwick undergraduates took part for payment. No participant had previously taken part in other experiments in this chapter.

Stimuli and Design. Three 5 × 5 checkerboard prototypes were created for each participant, using the same algorithm as in Experiment 14. One of the prototypes was used as the example feature in the instructions. The remaining 2 prototypes were used to construct the stimuli. These two prototypes will be referred to as elements a and b.

The checkerboard stimuli displayed to participants always measured 10 × 10 squares. Either no quadrants or one quadrants were replaced with one of the elements. In the initial stage of the experiment participants learned to categorize three types of stimuli, type A, type B, and other completely random checkerboards. Type A stimuli always included element a in one quadrant. The quadrant was selected at random on each trial, from one of the top two quadrants. The remaining squares of the checkerboard were randomly assigned to be black or white on each trial. Thus the only invariant part of the stimulus is element a. Similarly the only invariant part of type B is element b, which always appeared in one of the bottom two quadrants. The completely random checkerboards were included to ensure participants learned both types A and B.

In the transfer phase, new types, A' and B' were introduced. A' and B' differed from A and B , in that the element now appeared in the opposite half of the checkerboard. Where the element had appeared in the bottom half of the checkerboard, it now appeared in the top half, and vice versa.

Procedure. The procedure was identical to that of Experiment 14, with the intermediate phase omitted. In the first training phase participants classified type A, type B and random boards to the same criterion as in Experiment 14. In the transfer phase, participants classified 10 of each of types A, B, A', and B' and 10 random boards. They were specifically instructed to look out for elements in novel locations

in the transfer phase.

Results

Participants took between 3 and 8 blocks to reach criterion in the learning phase. Performance in the transfer phase was collapsed across elements. Thus, two conditions are compared: performance on checkerboards where the element appeared in a familiar location, and performance where the element appeared in a novel location. There was a significant difference between the two conditions, $t(6)=4.01$, $p<0.01$. Proportion correct in the familiar location condition (mean=0.80, S.E.=0.05) was higher than in the novel location condition (mean=0.39, S.E.=0.07).

Discussion

Participants were worse at classifying checkerboards when the diagnostic patch of a checkerboard appeared in a novel location. Thus it is possible that, as previously described, the results of Experiment 14 could be accounted for it terms of the location explanation rather than as feature creation.

Experiment 16

The results of Experiment 15 suggest that the novel location explanation may indeed account for the results of Experiment 14, without resort to a feature creation effect. The design and procedure of Experiment 16 are the same as Experiment 14. However, the stimuli are slightly different so that the effect of location is controlled across conditions. That is, for every new transfer stimulus, all elements appear in novel locations. Therefore this experiment could potentially demonstrate a feature creation effect that cannot be accounted for using the novel location explanation.

Method

Participants. 24 University of Warwick undergraduates took part for payment. No participants took part in other experiments in this chapter.

Stimuli and Design. Seven 5 × 5 checkerboard prototypes were created for each participant, as in Experiment 14. However the checkerboards participants were presented with were a different shape to that used in Experiment 14. They always measured 5 squares across by 16 squares down, and thus were columns of black and white squares. The elements appeared at some vertical location on the column, spanning the width of the entire column. In the initial stage of the experiment participants learned to categorize three types of stimuli, type A, type B, and other completely random checkerboards. Numbering the rows of squares from the top, Type A stimuli always included element a near the top of the checkerboard column, beginning at random at either row 2 or row 3. The small variation in location was introduced to reduce the likelihood that participants learned only a tiny part of a checkerboard. The remaining squares of the checkerboard were randomly assigned to be black or white on each trial. Thus the only invariant part of the stimulus is element a. Type B contained element b, near the bottom of the checkerboard, beginning at row 8 or 9. Completely random "other" checkerboards were included to ensure participants learned both types A and B. Note that throughout the experiment an element never appears right at the top of the checkerboard, or right at the bottom. It is assumed that if parts of an element that appeared at the edge of the board, then these would be easier to learn. This is because their appearance would remain constant, not be being altered by adjacent random squares, varying from trial to trial. If the two elements were swapped over to create the new stimuli for the transfer stage, the part that of the element that was at the edge of the board would now be in the center of the board, and would appear different in the different context. Performance on these items would therefore be worse. Although this effect would not produce an artificial feature creation result, as it would be equal across all three

conditions, the possibility of this effect was eliminated by keeping the top two and bottom two rows of every checkerboard random.

The checkerboards for the second training phase, type AB, type CD and type E-F, were created as follows. Type AB boards contain element a immediately above element b. The first row of element a in type AB began at either rows 2, 3 or 4. The remainder of the board was random, as previously described. Type CD boards contained the (previously unseen) elements c and d in the same arrangement, with the remainder being random. Type E-F boards contained elements e and f with e above f, but with two rows of random squares separating the elements, to ensure each element must be learned as a single feature. The first row of element e began at row 2, and the rows of element f began at row 9. The remaining squares were random. It was hypothesized that a configural feature, cd, would be learned to represent the conjunction of elements c and d. Learning of the analogous feature for the AB boards, ab, would be blocked by prior learning of features for a and b separately in the first part of the experiment. Separate features for elements e and f, would have to be learned as the elements are separated by a strip of random squares, varying on each presentation.

In the final stage of the experiment participants classified the previously described types AB, CD and E-F. In addition they classified three new types of stimuli, BA, DC, and FE. Checkerboard BA contained element b immediately above element a, with the remaining rows of the board being random. The rows of element a began at either rows 2, 3 or 4. Types DC and EF were similarly constructed. For all the new transfer stimuli elements appeared in novel locations. This was not true of all stimuli in Experiment 14, where only one transfer stimulus had elements appearing in a novel location.

Procedure. The procedure is the same as Experiment 14, except participants were warned that in the transfer phase the elements could appear in any location, and not just the locations they had seen them in, and they should lookout for this.

Results

The analysis of results here is the same as Experiment 14. Four participants did not complete the experiment within the one hour allowed, and are excluded from the following analysis. On average the remaining participants reached criterion after a mean of 4.0 and 3.5 blocks for the first and second learning phases respectively.

Of interest is performance in the transfer stage (Figure 40). For the new items a correct response is responding with the label of the type whose elements it contains. In all conditions, performance is worse on the new transfer items, replicating the location effect demonstrated in Experiment 14. The difference between the old checkerboard and the corresponding new checkerboard is largest in the compound condition, consistent with the hypothesis the feature creation effect.

As in Experiment 14, log odds scores were constructed for the three conditions. A one way ANOVA revealed a marginally significant difference in this log odds score between the three conditions, $F(2, 38)=2.80$, $p=0.073$ (Huynh-Feldt $\varepsilon=1.00$). t-tests showed that the log odds scores for the for the compound and blocked conditions differed significantly, $t(19)=2.11$, $p<0.05$. The difference between the log odds scores for the compound and elements conditions was marginally significant, $t(19)=1.90$, $p=0.073$. The marginal significance of these results suggests a replication of this experiment would be necessary before using this experiment alone as evidence for a feature creation effect.

Discussion

The results of Experiment 16 establish a feature creation effect that cannot be

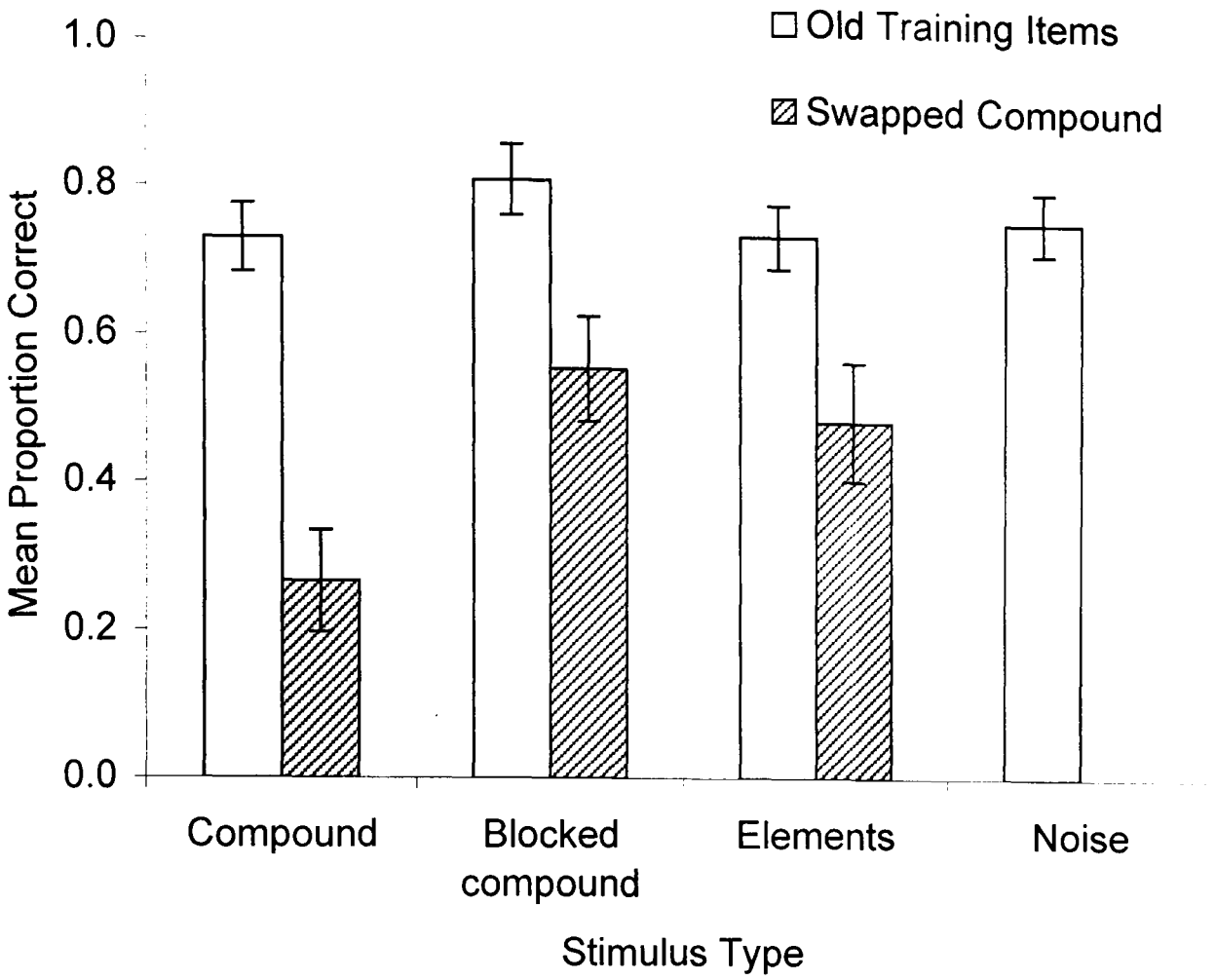


Figure 40. The mean proportion of correct responses for the transfer stimuli in Experiment 16. (Error bars are standard error of the mean.)

accounted for using the location hypothesis. For all of the new transfer stimuli, the elements appeared in novel locations. Participants categorized these stimuli less accurately than the corresponding stimuli with elements in familiar locations, consistent with the location hypothesis. However, the difference between classification accuracy for type DC stimuli and type CD stimuli (where participants were hypothesized to represent the stimulus using a single configural features for the conjunction of elements c and d) was greater than for type AB and type BA or type E-F and type FE (where participants were hypothesized to use two features – one for each element), consistent with the feature creation hypothesis. (This effect was much smaller than the effect in Experiment 14, which suggests that part of the effect in Experiment 14 was due to the location explanation.) Because in type CD stimuli elements c and d were always adjacent, with c above d, participants were hypothesized to form a single feature cd. This feature is not present in type DC, and so participants should not categorize type DC as type CD. Participants were hypothesized to have learned separate features for each of the elements in types E-F and AB. These features occurred in novel locations in the new transfer stimuli FE and BA. Thus although participants should be worse at classifying these stimuli, because they have been shown to be worse at identifying features in novel locations, they did show some tendency to categorize these stimuli accurately. This is because the features were present in the new transfer stimuli although they occurred in novel locations. In type E-F elements e and f are separated by some rows of random squares. Thus participants must learn two features, e and f, for each element. For type AB, which contained elements a and b adjacent to one another, participants already had learned the features separately when learning types A and B in the first stage of the experiment. Thus, learning of a single feature, ab, was blocked.

General Discussion

Experiment 9 replicated Schyns and Rodet's (1997) feature creation effect. Experiment 10 provides evidence that the effect may instead be due to participants' beliefs about the relative location of the features. Experiment 11 finds slight evidence to support this using an alternate stimulus set. Experiments 12 to 16 investigate feature creation using checkerboard stimuli. Together Experiments 12 and 13 provide evidence that features are created during a categorization task directly analogous to Schyns and Rodet's by showing that the learning of feature a single feature can be prevented by encouraging parsing of parts separately. Experiment 14 addresses some methodological problems in porting Schyns and Rodet's design to these new stimuli, and demonstrates a feature creation effect. Experiment 15 provides evidence for an alternate explanation, based on features being better recognized in familiar locations, but Experiment 16 demonstrates a feature creation effect that cannot be explained by this location based-account.

Similarity and The Creation of New Features

The concept of similarity pervades the psychological literature (Ananiadou, 2000). With perceptual stimuli, the implementation of similarity in models differs: some models use a spatial metaphor (e.g., Ashby & Perrin, 1988; Ashby & Townsend, 1986; Medin & Schaffer, 1978; Nosofsky, 1984; Nosofsky, 1986), and others use feature matching (Tversky, 1977; Tversky & Gati, 1982). The creation of new features will impact upon models using either implementation. In feature matching models, the creation of a new feature will alter similarity between items by (potentially at least) adding a new feature that may be compared between objects. In models that use a spatial metaphor, the creation of a new feature will transform the similarity space. For example, Palmeri and Nosofsky (in press) found that after

training participant on random distortions of checkerboards a multidimensional scaling recovery of participants' perceptual space revealed the prototypical checkerboards to be extreme points in psychological space. The idea that these checkerboards were extreme points before training seems unfeasible as, given the random nature of the stimuli generated, that the three prototypes generated (out of a possible $2^{256}=1.2x10^{77}$) just happened to be extreme points seems very unlikely. The fact that the psychological space differed from the assumed physical space (where similarity is based on the number of squares in common, and prototypes were category central tendencies) allowed an exemplar model to explain the prototype effect – where performance is most accurate of the prototypes, rather than the distant examples of each category. Note that selective/attentional weighting of dimensions (Nosofsky, 1986) is not sufficient to describe these changes in perceptual space. That the learning of features, and the resulting reorganization of perceptual space, allows exemplar models to predict a prototype effect that has been considered by some to be problematic for the models (see McLaren, Bennett, Guttman-Nahir, Kim, & Mackintosh, 1995; but see also Lamberts, 1996) and demonstrates the importance of considering the creation of features.

Biases in Feature Creation

When confronted with a set of novel stimuli participants could potentially learn any of a large set of possible features. The choice of features is important because it determines the usefulness of these features when encoding subsequent stimuli. Choosing features that describe the stimuli very accurately will lead to choosing features that describe the noise in the particular examples available to participants. Alternatively, if the choice of possible features is too greatly restricted, adequate features may not be available from the restricted set of possible features.

The problem of restricting the set of possible features that could be learned to prevent learning noise in current examples (allowing generalization to new examples), but not so much as to prevent a successful representation of the stimuli forming, is called the bias-variance dilemma (Geman, Bienenstock, & Doursat, 1992). Put simply, participants must be biased against learning some types of feature to prevent them learning irrelevant features, but not so biased that they cannot learn useful new features. Certainly identification of bias in the creation of new features will be important in developing a model of feature creation. Relevant evidence is reviewed here.

Participants seem to require time to learn new features, allowing them to encounter a large enough set of training examples. In Schyns and Rodet's (1997) Martian cell experiments and also Schyns and Murphy's (1994) Martian boulder experiments participants learned features after the order of ten 2 s presentations of each stimulus. In the checkerboard experiments (Experiments 4 to 8) presented here participants took of the order of 100 presentations of similar duration. In Goldstone's (2000) curved line segment experiment participants took slightly more trials, approximately 400, over which they gradually learned features (as demonstrated by gradually reducing classification latencies). Shiffrin and Lightfoot's (1997) participants took approximately 20,000 trials, again gradually learning features of the character made from spatially separate straight line segments (as measured by reduction in search slopes in a visual search task). Requiring more than one presentation of a stimulus to learn features is certainly a sensible strategy, as with only one presentation participants could not deduce which parts of the stimulus are invariant, and which parts will vary from exposure to exposure. But why is there such a large variation in learning times? I suggest that this is due to interaction with

other biases. As Goldstone points out, one such bias may be for learning of features to represent spatially contiguous units (Palmer, 1992; Palmer & Rock, 1994). In Schyns and Rodet's experiment the features were already segmented from the rest of the stimulus, appearing as cell bodies amongst spatially separate random cell bodies. In the checkerboard experiments in this chapter (and also, to some extent, in Goldstone's experiments) the diagnostic invariant feature needed to be segmented from the rest of the stimulus. Although in Shiffrin and Lightfoot's experiment the line segments were already segmented, it was only possible to create diagnostic features integrating several spatially separate segments. Goldstone compared two conditions that differed on whether the curved line segments to be integrated were connected or spatially separate (see also Czerwinski, Lightfoot, & Shiffrin, 1992). Although participants' categorization latency reduced with practice in both conditions the reduction was greater in the connected condition. Thus it seems that participants may be biased towards learning spatially separate units, although this bias can be overcome at the expense of additional training examples. An important caveat to this is that stimuli to be unitized must all be viewable in a single fixation. Goldstone demonstrated that stimuli that were enlarged so that they could not be taken in a single fixation showed no evidence of being unitized.

It may be that participants only learn new features when current features are not adequate for some new task. This is certainly consistent with the blocking demonstrated in the checkerboard experiments in this chapter, and in the experiments of Schyns and colleagues (Schyns & Murphy, 1994; Schyns & Rodet, 1997), where the existence of previously learned features that can be used in a new categorization prevents the creation of new features that would be learned in the absence of the previously learned features. Almost certainly, with prolonged practice, this blocking

can be overcome (Schyns & Murphy, 1994).

Another important factor may be whether participants are consciously trying to form new features. In Schyns and Rodet's (1997) experiment, and the experiments in this chapter, participants were instructed to look for features. Pilot work using the checkerboard stimuli demonstrated that few participants showed any learning of the categories without instructions on how to spot the features. In other experiments (Goldstone, 2000; Shiffrin & Lightfoot, 1997) participants were not instructed to try to learn features, but also took much longer to show evidence of learned features. However, the absence of explicit instructions does not exclude the possibility consciously trying to learn new features. In Goldstone's experiment, participants gave a self report of the strategy they felt they had used. In the condition where line segments were required to be unitized, 75% of participants reported trying to remember an "overall image" of the stimulus, against only between 7% and 25% in the other condition where unitization was not required or beneficial. I do not wish to claim that conscious effort is always required for learning new features – this would certainly be at odds with a large body of evidence in the perceptual learning literature – just to suggest that explicit instruction and conscious effort can greatly speed the process. Consistent with this is an example from perceptual learning. Ahissar (1999; Ahissar & Hochstein, 1997) trained participants on the detection of odd orientation line segment in an array of line segments under very brief exposure. After over a thousand trials most participants' detection was not significantly above chance accuracy. However another set of participants who received only one single easy trial, where the odd orientation differed greatly from the orientation of the distractors, performed much more accurately, and significantly above chance, on the more difficult discrimination.

Properties of New Features

Although features speed the processing of stimuli (Goldstone, 2000; Shiffrin & Lightfoot, 1997), and qualitatively alter the perception of stimuli as demonstrated in the experiments presented here (see also Schyns & Rodet, 1997) little is known about the other properties of features. One possibility is that with practice features become like simple primitive visual features, with properties such as attentional capture. Certainly the claim that features were only cognitive constructs (see Schyns et al., 1998 for an overview) can be extended because of the evidence that learned features can dramatically reduce categorization latency (Goldstone, 2000) and search slopes in visual search (Shiffrin & Lightfoot, 1997). For features to have an impact in these perceptual tasks, significantly reducing detection time, it seems that there are either large top down processes from cognition to perception facilitating detection of the features, or that the locus of learning is perceptual rather than cognitive.

Conclusion

There is mounting evidence that new features may be created to meet the requirements of new tasks, facilitating perception. Although an experiment in this chapter indicates an alternative explanation for Schyns and Rodet's (1997) finding that new features may be created, the other experiments demonstrate this result with a new class of stimuli.

Chapter 5

General Discussion

Summary

The Effect of Category Variability in Perceptual Categorization

Two very different views have been advanced in the categorization literature concerning how people learn categories from labeled examples. The exemplar view suggests that people store some or all examples, and categorize new items by their similarity to these stored items. What we call the distributional view suggests, instead, that people fit probability distributions using the examples from each category, and classify new items by reference to these probability distributions. A key differential prediction between these viewpoints concerns the classification of new examples precisely intermediate between the nearest examples from two categories that differ in variability. The exemplar approach, illustrated using the generalized context model (Nosofsky, 1986), predicts that the intermediate item should typically be classified in the lower variability category. This is because the examples of the low variability category are clustered nearer in perceptual space to the new examples, and are therefore more similar. By contrast the distributional approach, illustrated using general recognition theory (Ashby & Townsend, 1986), predicts classification into the higher variability category, as these new examples are more likely to belong to the high variability category. The experiments in Chapter 2 investigated classification of items intermediate between two categories differing in variability. Neither prediction was confirmed experimentally – instead a highly variable pattern of results was found. Experiments 1 and 2 showed that classification behavior can be strongly influenced by the salience of the difference in variability between the categories. The greater the salience of the variability, the more likely participants are to classify the new examples into the higher variability category. Experiments 3 and 4 showed great variation between participants on the effect of

increasing the difference in variability between the two categories. When the difference in variability was increased, some participants increased the proportion of high variability responses to the intermediate items, and others showed a decrease. Neither the exemplar nor distributional viewpoints can predict the behavior of the majority of participants. This presents a serious challenge to the continuum of models of categorization between the exemplar and distributional viewpoints (e.g., back propagation and radial basis function models).

Identification and Categorization of Simple Perceptual Stimuli: A Memory and Contrast Strategy

Traditionally research rooted in categorization assumes that the cognitive system has access to an accurate representation of the absolute magnitudes of the properties (e.g., frequency, loudness, size) of the complex multidimensional stimuli that people routinely encounter (e.g., Ashby & Townsend, 1986; Nosofsky, 1986). However, research rooted in identification of simple perceptual stimuli suggests people have very poor representations of absolute magnitude information. People are poor at making absolute magnitude judgments, and are typically only able to divide stimuli into about five 'bins' on a single dimension (cf., Miller, 1956). In addition, judgments about absolute magnitude are strongly influenced by preceding material (e.g., Ward & Lockhead, 1971). The experiments in Chapter 3 investigated such sequence effects in categorization tasks. Strong sequence effects were found. Classification of a borderline stimulus between the two categories was more accurate when preceded by a distant member of the opposite category than a distant member of the same category. This category contrast effect is a serious challenge to existing models of classification – an exemplar model of categorization, adapted to predict sequence effects by assuming recently encountered stimuli are more influential in the

categorization decision, is constrained to predict the opposite pattern of results. A memory and contrast (MAC) strategy, where categorization is instead based on the relative perceived difference between the current stimulus and the preceding stimulus, is able to predict the pattern of results.

Feature Creation

Recent work in categorization (e.g., Goldstone, 2000; Schyns et al., 1998; Schyns & Rodet, 1997) and visual search (Shiffrin & Lightfoot, 1997) suggests that new visual features are constructed after experience with novel stimuli. These features facilitate the detection of the novel stimuli. Schyns and his colleagues (Schyns & Murphy, 1994; Schyns & Rodet, 1997) also demonstrate that new features qualitatively alter the perception of the stimuli. Different training orders were used to induce different sets of novel features in two groups of participants, with one group hypothesized to use a single to represent a stimulus, and the other hypothesized to use two features. This explanation is used to explain the two group's orthogonal categorizations of identical critical stimuli. The critical stimulus consists of two separate views of parts of the compound stimulus. The group with a single feature to represent the compound stimulus should not classify this stimulus into the same category, as the entire compound is never shown. However, the group with separate features for each part should, as both separate features are present in the critical stimulus. Replications of these experiments in Chapter 4 casts doubt on this claim. Addition of a simple background context to the test stimuli removed the difference in categorization between the two groups, with neither group classifying the critical stimulus into the same category as the compound stimulus. This is consistent with the alternate hypothesis that the two groups have the same feature set, but differ in their knowledge that the compound stimulus is made up from a

conjunction of the other stimuli. With the addition of the background context it is clear that the parts viewed on test are not joined to any other parts. and hence the that they cannot be part of the compound stimulus.

The remainder of the experiments in Chapter 4 are concerned with establishing a feature creation effect in a new class of checkerboard stimuli. Three methodological solutions to problems are provided in the development of a new paradigm: (a) participants do not have to remember features over long blocks of trials with similar stimuli; (b) a potential confound where by the feature creation effect obtained can be explained by features being better recognized in familiar locations is eliminated; (c) the avoidance of sequential presentation of the parts of the test stimuli is eliminated, allowing the measurement of reaction times. A feature creation effect is demonstrated that is consistent with participants learning the largest block of invariant squares in the checkerboard to represent stimuli. It is also shown that learning of a feature can be blocked by prior learning of features.

Category Contrast Effects in the Category Variability Experiments

Given the demonstration of sequence effects with simple geometric stimuli in Chapter 3, it seems likely that the category variability experiments of Chapter 2, which also used simple visual stimuli, would also show sequence effects. A meta-analysis of the data from the experiments in Chapter 2 was run to explore the possibility of category contrast effects. In the experiments in Chapter 2, the trials were in a random order. The critical pairs were not over-represented as they were in the experiments of Chapter 3. As we are only interested in a small proportion of possible pairs of trials, a large amount of data is necessary to investigate these effects. The two circles with dots experiments (Experiments 1 and 2) had a very small number of trials, and so do not provide enough data for the analysis. However,

the rectangles and ellipses experiments (Experiments 3 and 4) do contain a sufficient number of trials, provided stimuli are collapsed into groups. For each category the ten stimuli were collapsed into three groups: near, intermediate and distant. The three stimuli nearest the category boundary were in the near group, the three furthest from the boundary were in the distant group, and the remaining items in the intermediate group. With such a grouping, there is enough data to consider the effect on classification of a near stimulus of the immediately preceding stimulus. Specifically, the size of the category contrast effect may be determined by comparing classification accuracy of a near stimulus when preceded by a distant stimulus from either the same category, or the other category. Figure 41 shows the category contrast effect for Experiment 3. The mean proportion of correct responses to near stimuli on trial n that were preceded by a distant stimulus on trial n-1 was plotted as a function of whether the distant stimulus came from the same category or the other category. As the category structures were not symmetrical, jumps from the low variability category towards the high variability category (low to high jumps) are plotted separately from jumps from the high variability category towards the low variability category (high to low jumps). Figure 42 shows the analogous plot for Experiment 4. A category contrast effect, where performance is better when the distant stimulus came from the other category, is evident in the data from both experiments, for both types of jumps. For both experiments, high to low jumps give a bigger category contrast effect than low to high jumps. This description of the results in confirmed by a 2×2 ANOVA (category of tone on trial n-1 $\times$ jump direction). For Experiment 3 there was a main effect of category of tone on trial n-1, $F(1, 31)=9.61$, $p<0.005$. There was no main effect of jump direction, $F(1, 31)=1.15$, $p>0.05$. The interaction was significant, $F(1, 31)=5.83$, $p<0.05$. For Experiment 4

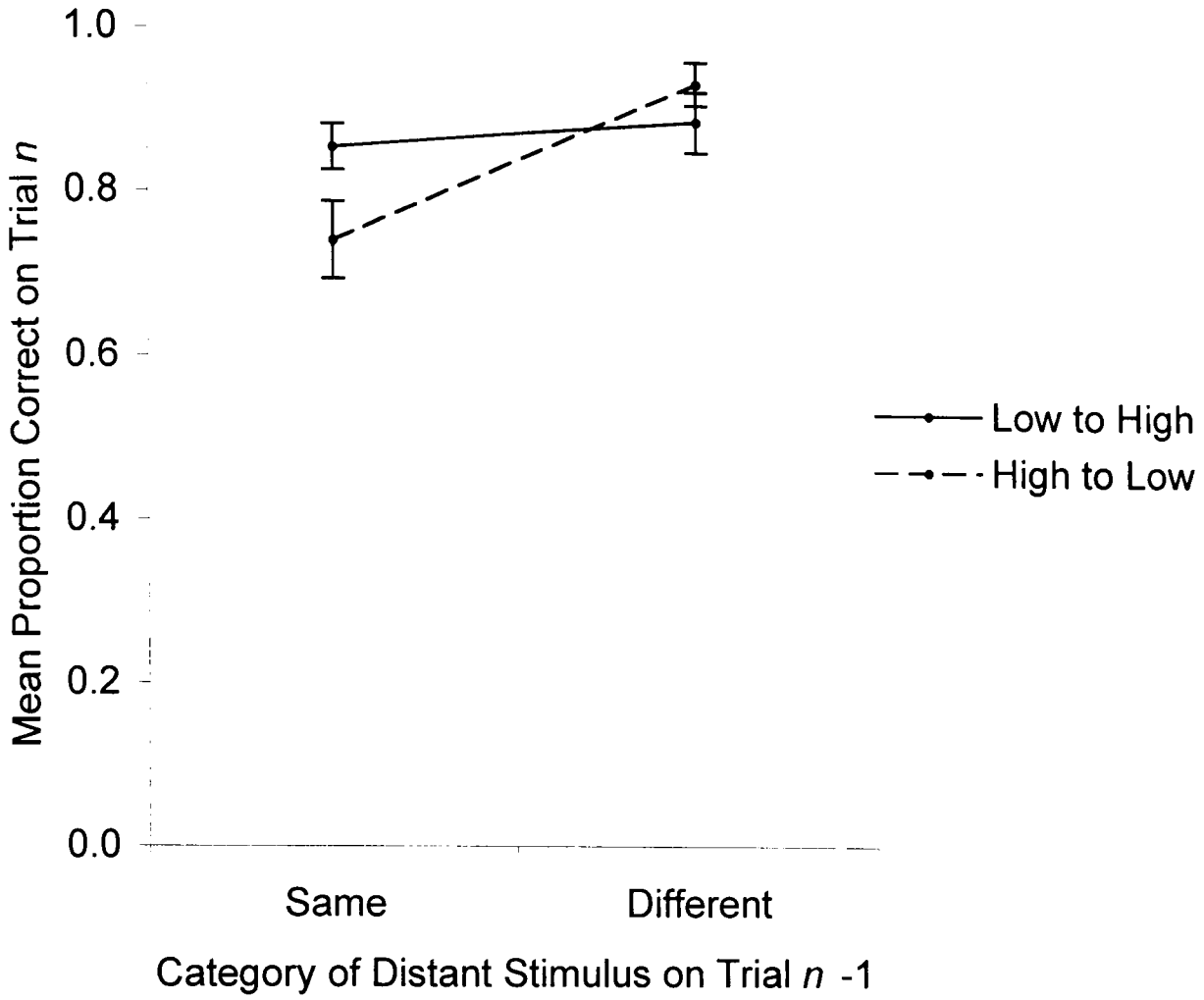


Figure 41. The mean proportion of correct responses on trial n as a function of the whether the distant stimulus on trial n-1 came from the same category or the opposite category for Experiment 3. The two lines correspond to trial pairs where there was either a jump from the low variability category towards the high variability category, or vice versa.

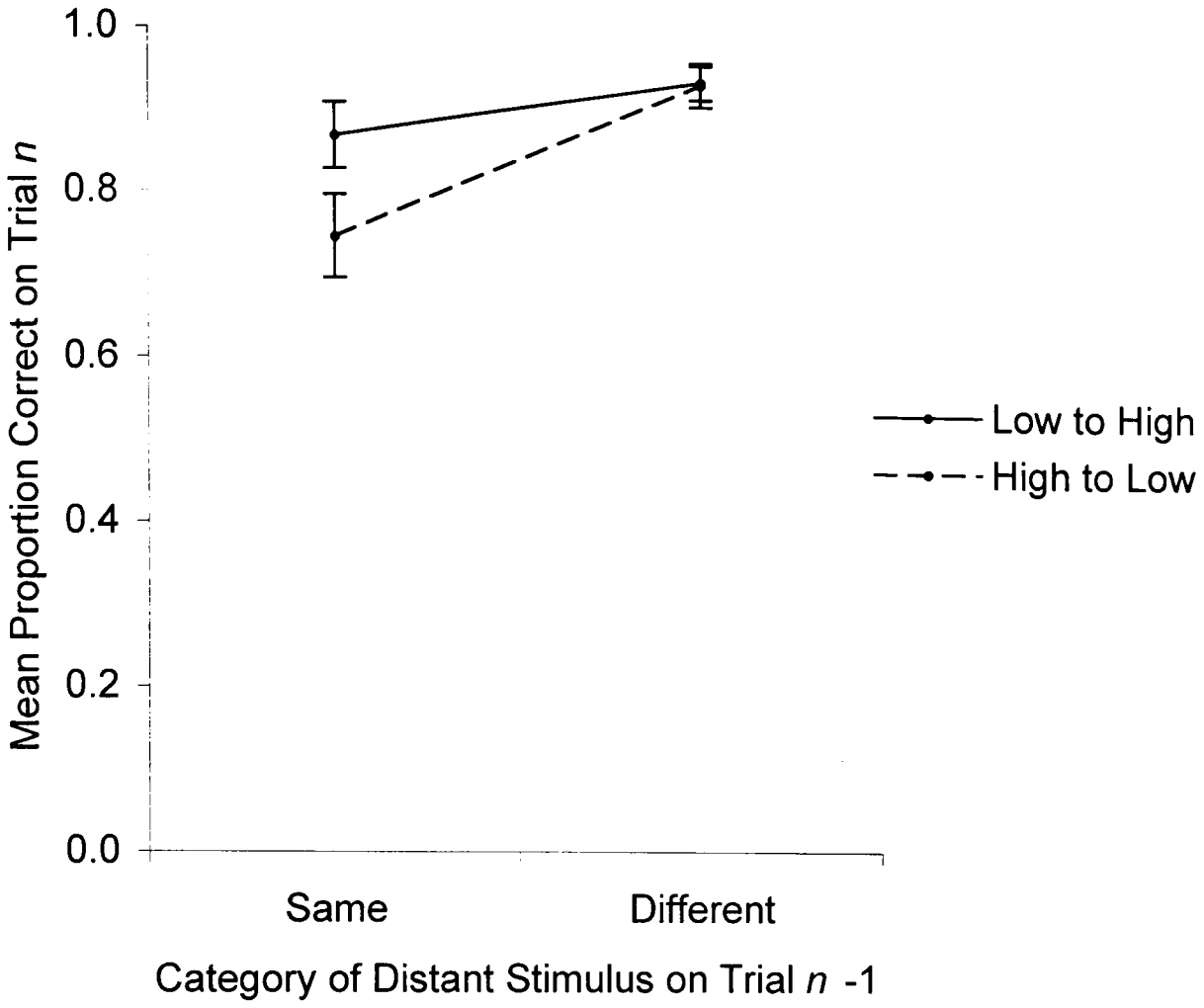


Figure 42. The mean proportion of correct responses on trial n as a function of the whether the distant stimulus on trial n-1 came from the same category or the opposite category for Experiment 4. The two lines correspond to trial pairs where there was either a jump from the low variability category towards the high variability category, or vice versa.

there was also a main effect of category of tone on trial n-1, $F(1, 31)=11.35$, $p<0.005$. There was no main effect of jump direction, $F(1, 31)=2.86$, $p>0.05$, although the effect was approaching significance. The interaction was not significant, $F(1, 31)=2.66$, $p>0.05$, but was approaching significance.

The category contrast effect is smaller than was observed with the tones experiments in Chapter 3 (Experiments 6 and 8), and approximately the same size as observed in the Nosofsky stimuli experiment (Experiment 7). A MAC strategy account of this difference in effect size is given later. The low to high jumps yielded a smaller category contrast effect than the high to low jumps. A MAC strategy predicts this result. Consider the category structure illustrated in Figure 43. The largest within low variability category jump (between items 1 and 5) is smaller than the smallest between category jump (between items 5 and 6). However, many of the possible within high variability category jumps (e.g., between items 10 and 6) are larger than the smallest between category jump. Thus the category contrast effect measured using jumps from the low variability category towards the high variability category (i.e., 1→5 compared to 1→6 – up jumps) will be zero, as there is never any confusion over whether one has jumped up the scale far enough to be in a new category. However, the category contrast effect measured from the high variability category towards the low variability category (i.e., 10→6 and 10→5 – down jumps) will be larger, as the large within category jump 10→6 could be confused with a between category jump. This is true even if two c parameters are used in Equation 23, one for up jumps and one for down jumps. Numerical modeling (not presented here) using the MAC model detailed in Chapter 3 confirms this argument. for both the one and two c parameter models.

The existence of category contrast effects in the data from the category

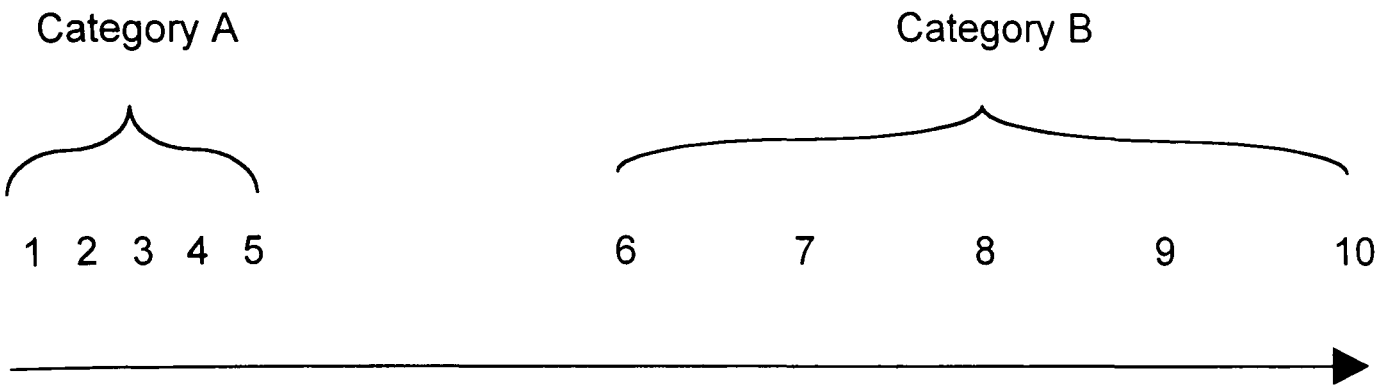


Figure 43. Two categories separated where one category is more variable than the other.

variability experiments raises the question of whether the MAC strategy can account for the effects of category variability, and change in the relative variability, on the classification of items intermediate between the two categories.

A Memory and Contrast Account of Category Variability Effects

The MAC strategy can predict the category variability effects observed in Chapter 2 with a slight extension of the model. Separate $\underline{c}$ parameters for jumps up the scale and jumps down the scale need to be introduced, rather than one parameter for jumps in both directions. This addition of this free parameter can be justified on two counts. First, jumps up a dimension seems qualitatively different from jumps down the scale. It seems very likely that participants can always tell which direction a jump is in (provided it is sufficiently large, as it is in the category contrast experiments). Second, with asymmetrical categories the optimum $\underline{c}$ parameter for jumps up the scale is different to the $\underline{c}$ parameter for jumps down the scale. This is because the relative sizes of the within category and between category jumps differs for jumps up the scale and down the scale.

One further observation need be made before the MAC explanation of the variability effect is described: Small deviations from the optimal $\underline{c}$ parameters hardly reduce overall classification accuracy in a random sequence of examples. Thus participants may be inaccurate in choosing the optimal parameters. Therefore, it is likely that under such circumstances each participant may use different pair of $\underline{c}$ parameters. Figure 44 shows the probability of classifying items into the low variability category as a function of the item's position along a single dimension, as predicted by the MAC strategy. The category structure used is that illustrated in Figure 43. (Specifically, items 1 to 10 had values 1, 2, 3, 4, 5, 10, 12, 14, 16 and 18 on the single dimension.) The overall classification of a given item is determined by

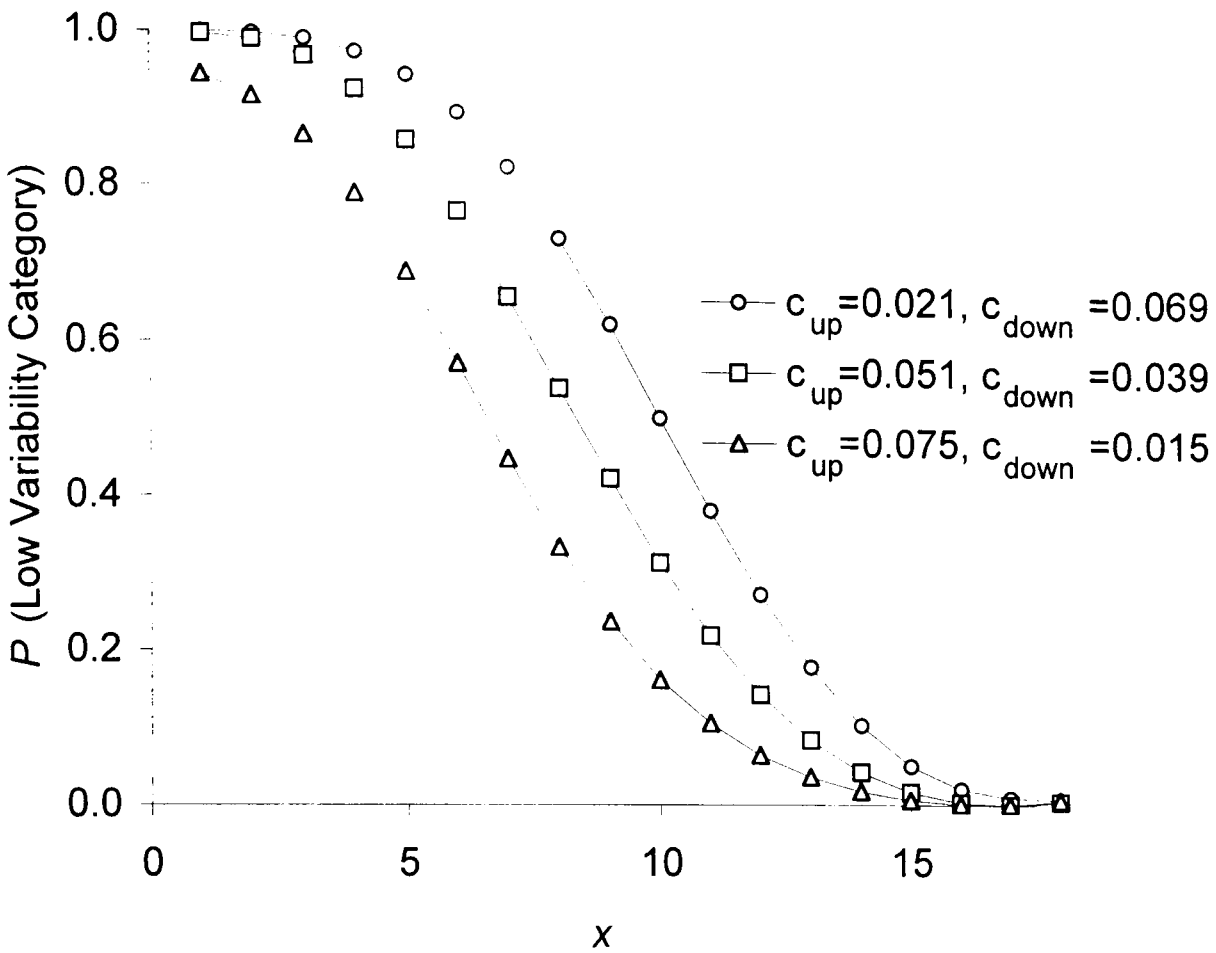


Figure 44. The probability of a correct response as a function of the value of the stimulus on the single dimension. The three lines correspond to the predictions of the MAC model with three different pairs of c parameters.

taking the average of all the probabilities of classification of the item into the low variability category when the item is preceded by each possible training stimulus. The three lines on Figure 44 correspond to three different values of pairs of c parameters. The central line corresponds to the optimal values of the two c parameters, with an overall average accuracy of 91.5% on training items. The lower line corresponds to an overall accuracy of 89.2%, and the upper line to an accuracy of 89.3%. Small deviations from the optimal values of the c parameters produce only very small accuracy reductions, but show a large change in the proportion of items intermediate between the two categories classified into the low variability category (or into the high variability category). Thus the individual variations in the sensitivity to variability are accounted for. If it is assumed that participants select different c parameters for the 1:2 and 1:4 (or 1:2 expanded) category pairs in Experiment 3 (and 4), then the change in the proportion of transfer items classified into either category can also be explained. This assumption is not unlikely, given the change in the stimuli between the two conditions.

In summary, the MAC can strategy predict the category variability effects observed in the experiments in Chapter 2. However, the size of the category contrast effect observed in the Chapter 2 experiments is small, approximately four times smaller than predicted by the simple MAC strategy. This suggests that a MAC strategy is not the sole strategy used by participants. If the MAC strategy is only making a small contribution to participants' performance, as is evident form the small category contrast effect, then it cannot be the entire explanation for the variability effects. However, if information from comparison with stimuli from trials further back in the sequence (i.e., n-2, n-3, … and so on) is being used a MAC strategy predicts a much smaller category contrast effect. A borderline stimulus

preceded by a distal stimulus on trial n-1 is unlikely to have been preceded by another borderline stimulus on trial n-2, and therefore the stimulus on trial n-2 is likely to reduce the error induced by the distant stimulus on trial n-1, reducing the size of the category contrast effect. If one simply averages the probability of responding with a given category label deduced from the jump between trial n and trial n-1, with the corresponding probability with the jump between n and n-2, then the category contrast effect is approximately halved in magnitude. Figure 45 shows the category contrast effects predicted for the category structure shown in Figure 43 using the MAC strategy with a single c parameter (similar predictions are made by the two parameter model). The top panel shows the large category contrast effect predicted using only the jump size from trial n-1 to trial n, (for optimal c=0.045). A larger category contrast is predicted for jumps down the scale category, as described above. The bottom panel shows the corresponding effects when the jump size from trial n-2 to trial n is also used. If the correct classification cannot be deduced from either jump size, the probabilities for classification into the low variability category are calculated using both jumps, and then averaged together. (Cases when the classification can be deduced are those when the stimulus on trial n is a more extreme category member than a stimulus from the same category on trial n-1 or trial n-2.) Thus the MAC strategy can predict smaller category contrast effects. Further, this adaptation to use extra information from further back in the sequence does not alter the fact that changing the c parameters for the up and down jumps (within the range that allows high accuracy) will alter the position of the generalization gradient.

Extension of the Mac Model to More Distant Trials

In the above discussion it was suggested that a MAC strategy may be extended to make use of relative magnitude information deduced from trials further

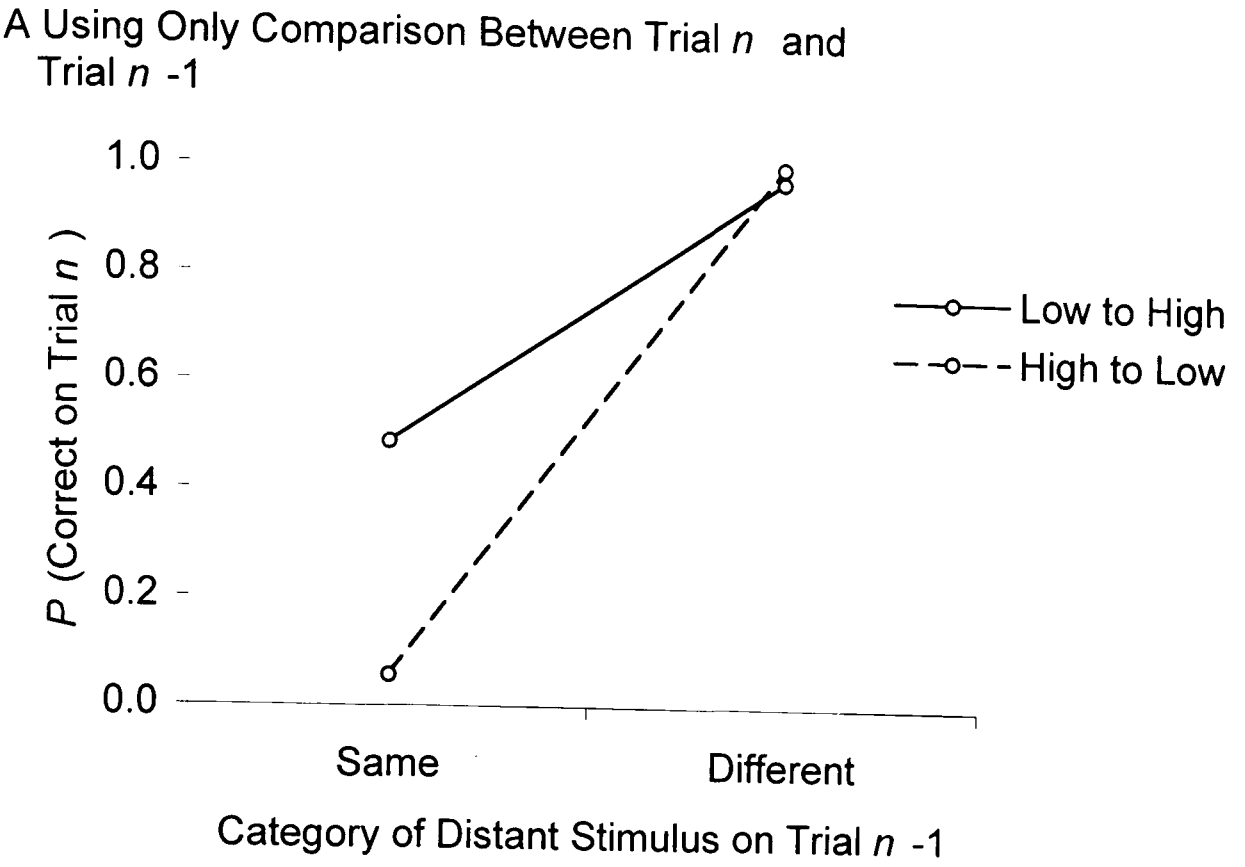


Figure 45. The mean probability of a correct response on trial n as a function of the whether the distant stimulus on trial n-1 came from the same category or the opposite category. The two lines correspond to trial pairs where there was either a jump from the low variability category towards the high variability category or vice versa.

back than trial n-1. Three questions need to be addressed when considering this extension. First, how the availability of the relative magnitude information from the comparison of trial n and trial n-x changes with x. As x increases presumably the relative magnitude information is less available – forgetting. Secondly, whether the use of information from trial n-x, changes as a function of the intervening material (trials n-x+1, n-x+2, …, n-1). Mori's (1989) demonstration that in absolute identification the use of information from trial n-1 increases as the availability of information of trial n is decreased suggests that the use of information may well depend on the intervening material. The final question is how the information from each of the preceding trials is combined to maximize the likelihood of a correct response. These second and third questions may be related – the method of combining the information may well depend on the intervening material.

The answer to the first question, of availability of information from previous trials, is an empirical question. Such availability can be measured using a simple discrimination task, where participants are asked to make same / different judgments for the first and last items in a sequence. By varying the number of intervening items, a measure of the availability about the relative magnitude of the difference between the stimuli can be obtained. It is possible that this availability will depend on the nature of the stimuli.

In considering the answers to the last two questions posed above, in the absence of empirical data, let us assume that information is available equally from trial n-1 and trial n-2. This is almost certainly not the case, but such an assumption makes consideration of the last two questions simpler. Here an extension of the model to include information from trial n-2 is given. In a categorization task where feedback is given, the participants' task is to categorize the stimulus given on trial n,

given knowledge of the categorizations on trial n-1 and trial n-2, and the jump sizes between the these previous trials and the current trial. If participants do indeed use information from trial n-2 in addition to information from trial n-1, it is not clear what strategy participants might use to combine the two sources of information. In the simple discussion earlier in this chapter, two probabilities of responding with a given category label were calculated, one from the n-2 to trial n jump size and the other from n-1 to trial n jump size. These two probabilities were then averaged together. This is certainly not the optimal strategy, as it leads to only a slight increase in overall accuracy. On trials where the n-1 to n jump size leads to a low accuracy of a categorization of the stimulus on trial n, averaging of the probability with the probability from calculated from the n-2 to n jump leads, averaged across all possible stimuli on trial n-2, to a reduction in error. However, for instances where the n-1 to n jump size leads to a high probability of a correct categorization, averaging of the probability with the corresponding probability from the n-2 to n jump leads, on average, to an increase in error. This is because most of the time the stimulus on trial n-2 is not the same as the stimulus on trial n-1. The slight increase in overall accuracy is only due to the fact that sometimes the jump between the stimulus on trial n-2 and trial n leaves no uncertainty as about the correct categorization, when the stimulus on trial n-1 does. For example, consider the following sequence of stimuli from the category structure in Figure 20: 5→2→4. Using the jump of +2 between trial n-1 and trial n the categorization of the stimulus on trial n is not determined. The jump may or may not correspond to a crossing of the category boundary. However if the trial n-2 to trial n jump size of -1 is considered, given that it is known that stimulus 5 belongs to category A, and category A members are low in value on the stimulus dimension, then a stimulus one unit lower must also be also

belong to category A. Hence the categorization of the stimulus on trial n, which was uncertain if only information from trial n-1 was used, is no longer uncertain.

As the two sources of information are conditionally independent given n, i.e., knowledge of the trial n-1 to trial n jump size does not allow prediction of the trial n-2 to trial n jump size and vice versa when the value of the stimulus on trial n is known, then the optimal strategy is to combine log probabilities. However, the discussion of the combination of the two probabilities is slightly academic for two reasons. First, it is unlikely that the perceptions of all sizes of jumps are equally reliable. The reliability of the perception of different jump sizes is an empirical question. It is already demonstrated that large jump sizes in absolute identification show a greater degree of assimilation (e.g., Ward & Lockhead, 1970; Chapter 3). Second, it is possible that there is interaction in the perception of jump sizes, with the size of the jump between trial n-1 and trial n affecting perception of the jump size between trial n-2 and trial n. In other words, the perception of the two jump sizes may not be independent.

Feature Creation and MAC

Evidence in Chapter 4 suggests that experience with novel stimuli can lead to the creation of new features. If one conceives of stimuli as represented in a psychological space, then this would lead to stimuli moving substantially about in the psychological space. However, it is not clear how a MAC account might sit with the creation of new features, because in a MAC account this perceptual space representation is not assumed.

When categorization is based on presence or absence of a feature, there is no need to resort to a MAC strategy. For example, when deciding whether a vivid color red is present or absent in a stimulus composed otherwise of vivid blue and vivid

green is easy. When categorization is based on the degree to which a feature is present, participants may need to resort to a MAC strategy. For example, consider a set of stimuli made by interpolating between two novel stimuli, such as the Bezzier curves used by Goldstone (1994). Goldstone (2000) demonstrates that new features are created to after exposure to similar stimuli. Although it remains an open empirical question as to whether participants categorizing these stimuli into two groups will show a category contrast effects it seems likely that they might.

Goldstone (1994) demonstrated that in a classification task participants show improved discriminability of a pair of adjacent stimuli that fall across the category boundary compared to a control pair within the category. This leads to the intriguing possibility that participants create new features to improve discriminability. Before the creation of the new features, classification was based on the degree of many existing features that varied continuously between stimuli. In the absence of absolute magnitude information, participants would be forced to use a MAC strategy. New features may have been created, that would allow the categorization to be reframed in terms of the presence or absence of these new features, and therefore the task which would no longer require a MAC strategy. Such a shift in categorization strategy was not observable in the experiments in Chapter 4 as the creation of new features was not possible, given the simple one dimensional nature of the stimuli.

The empirical test of this possible change in strategy would be to measure simultaneously the category contrast effect and the reduction in classification latency and increase in discriminability consistent with the creation of new features. In more detail, the design of the experiment would be as follows. Create a novel dimension by taking two novel features, and then create a set of stimuli morphed between the two features. Expose participants to this stimulus set in a distractor task where they

do not make categorizations to allow new features to be learned to represent the stimuli. Then in a categorization task with the same stimuli observe the category contrast effect, the reduction in categorization latency, and the improvement in discriminability on the category bound (relative to equivalent within category discriminations). Initially a large category contrast effect should be observed. If additional new features are created to facilitate the categorization, there will be a reduction in categorization latency, and an increase in discriminability on the category bound, together with a reduction in the category contrast effect. These changes would correspond to a transition from a MAC strategy, to a simple classification based on the presence or absence of new features.

In summary, the hypothesis is that while participants may initially have to rely upon a MAC strategy to classify stimuli varying along a psychological continuum, new features may be created to facilitate categorization, provided the stimuli are of sufficiently high dimensionality. This hypothesis is certainly consistent with the extensive reanalysis of the categorization literature for the most frequently used category structure conducted by Smith and Minda (2000). They demonstrated that provided prototype models could reproduced an accuracy advantage for frequently seen items, the models fit the data as well as exemplar models. A MAC model will produce a very similar pattern of responding to a prototype model when predicting averages over all preceding stimuli, and the creation of new features for frequently viewed items would provide the improved accuracy for old training items.

Summary of Chapter

In the preceding discussion links between the three experimental programs in this thesis were described. It was shown that a MAC strategy is able to predict the category variability effects. In categorization of more complex stimuli, it was

suggested that reliance upon a MAC strategy may be replaced with use of new features created to reframe the categorization task in terms of the presence or absence of features, rather than in terms of the degree of presence of features. Related experimental work was suggested to investigate these hypotheses. The final sections of this chapter are further suggestions for extension of the work presented in this thesis based on submitted grant proposals.

Further Work: Memory and Contrast in Identification and Categorization

The following experiments described are designed to inform the extension of the simple MAC model described in Chapter 3. First, the basic category contrast effect could be replicated and the generality and magnitude of the effect explored, using the methodology from Chapter 3. Extension of this effect to other dimensions (e.g., loudness, line length, brightness, orientation) will establish its generality. Comparison of the magnitude of the effect across experiments will indicate the relative reliance upon relative magnitude information, or alternatively, how the availability of relative magnitude over time changes between stimuli. The use of categories constructed from multidimensional stimuli will further explore the generality of the effect. Experiment 7 has demonstrated the category contrast effect for multidimensional stimuli used by exemplar theorists (semi-circles of varying radius, with radii of varying orientation: Nosofsky, 1986). In this demonstration both dimensions were correlated, separable, and diagnostic of category. With multidimensional stimuli, the size of the difference between stimuli on one dimension at which it is optimal to switch categorization response can be made to depend in carefully controlled ways on the value of the stimuli on a second dimension. The degree to which participants are able to use information on the second dimension is predicted to depend on their ability to perceive absolute

magnitude or infer it from experimental context. This work will enable the construction of a new theory of representation of multidimensional stimuli based on relative magnitude information.

Through analysis of the effects of previous stimulus/response pairs on a current response using multivariate information transmission (McGill, 1954), the relative dependency of the current trial on previous trials can be determined. Such an approach has been successfully used in absolute identification (Garner, 1953; McGill, 1957; Mori, 1989). The central idea is that inter-trial dependencies over various spans (e.g., between trial n and trial n-k) can be controlled in such a way that information transmission between trial n-k and trial n may be computed directly. Such analysis will reveal any use of information other than that from immediately preceding trials in the categorization decision.

Previous work has not separated out effects of intervening trials from effects of intervening time. In separate experiments, manipulation of the inter-trial intervals will allow observation of how between trial dependencies vary as a function of the availability of perceived differences between trials. One such manipulation is illustrated in Figure 46. In the even spacing condition, greater effect of more recent trials is predicted on trial n, i.e., the amount of information transferred from trial n-1 to trial n should be greater than the amount of information transferred from trial n-*2* to trial n, and so on. When the difference between the availability of previous trials is smaller, the amount of information transferred should be less disparate. The results of experiments will allow direct comparison of these categorization studies to the large literature on temporal memory and forgetting.

Measurement of the sequential effects observed in absolute identification, and (within the same experiment) the category contrast effect, will allow the effect of

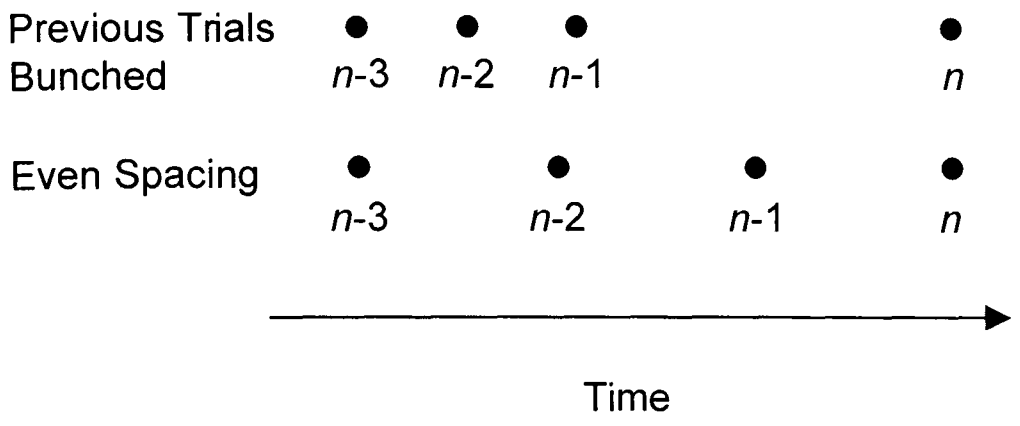


Figure 46. Two conditions where the spacing of trials varies.

previous stimuli on the perceived magnitude of the current stimulus to be factored out. After estimates of the subjective magnitude of the current stimulus have been determined in an absolute identification task, existing models of categorization, such as the GCM (Nosofsky, 1986), will be applied using these derived subjective magnitudes as input. In other words, the biased estimates of absolute magnitudes observed in an absolute identification task can be used to provide an independent assessment of the data participants have available to them in a categorization decision. Such data will be vital in constraining the developing unified account of identification and categorization. Appropriate data will be obtained from a small number of participants during extended sessions of alternating absolute identification and categorization. Such a procedure was used in Experiment 8, and will be extended to other stimulus sets.

Until now only stimulus sets that are evenly distributed along a dimension have been considered (with the exception of the stimulus structure in the category variability experiments). The MAC account predicts altering of the spacing of the stimulus set along a dimension, so that, for example, if one category's inter-stimulus spacing is larger than the other categorization performance will be affected substantially (because the relation between ITD and category-shift probability will become asymmetrical). Data on the ability of participants to deal with categories differing in variability will be important in informing modeling on the categorization process.

Performance on pseudo-random trial sequences where the relative frequencies of the magnitude of the change in stimuli between adjacent trials are manipulated could be manipulated in a series of experiments. The MAC model predicts that if participants are sensitive to jump sizes, and determine an optimal

jump size for changing responding, then such a manipulation may lead to erroneous estimation of the correct jump size. This would be the case if general properties of the sequence effects act as initial learning biases, as is the case for assumptions of category distribution (Flannagan et al., 1986). Evidence for the use of such expectations would provide vital constraints on the modeling.

The final suggestions for further experimental work to explore the MAC account concern manipulation of the availability of absolute magnitude information. A plausible working hypothesis is that reducing the certainty of absolute magnitude information would encourage use of a MAC strategy. Manipulations in luminance and duration of visual stimuli and signal to noise ratio of auditory stimuli have been successfully used in absolute identification to manipulate the relative inter trial sequential dependencies previously discussed (McGill, 1957; Mori, 1989). Similar methods could be adopted to vary the predicted necessity of MAC-like strategy usage. The withholding of feedback in alternating blocks could also be used to reduce the availability of absolute magnitude information. Such a manipulation is known to alter sequential dependencies in absolute identification (Mori & Ward, 1995); according to the MAC model this will lead to predictable effects on categorization performance. Multidimensional stimuli may also be used to investigate this issue. With a stimulus structure where the absolute magnitude on one dimension is needed to select the optimal jump size on a second dimension should reduce the effectiveness of MAC strategies, and therefore we predict smaller category contrast effects (and probably poorer average accuracy).

The proposed experimental work would inform the development of the MAC model in two ways: (a) Inclusion of information from more than one previous trial in the decision process, as described above. The relative weighting of trials would be

provided by data from the proposed experiments. Links with existing models of memory will be established using the data from the varying inter-trial interval experiments. The assumption that responses are not independent can be examined by seeing whether the response sequences violate the assumptions of a Markov model (Feller, 1950). (b) The MAC model and standard exemplar models can be conceptualized as lying at opposite ends of a continuum. At one end the MAC model presented here embodies the hypothesis that the decision process is based entirely on the difference between the current and the immediately previous trial together with knowledge of the category label associated with the previous trial. At the other end the exemplar model says the decision process is based on the jump size between every previous example and the current example, together with category labels for each previous exemplar. The development of a hybrid model would be useful in describing data obtained from the proposed experiments. Should the information limits of the response process be revealed in categorization by the studies, then a hybrid model must be adapted to explain how information selected varies as a function of information available.

Further Work: Feature Creation for Novel Visual Stimuli

The experiments proposed here focus on two key questions concerning how these novel features may function: (a) Can features attract attention to themselves, facilitating their detection? (b) In building a representation of a stimulus, can features of the same stimulus be processed in parallel? Existing evidence relevant to the first question is reviewed in the General Discussion of Chapter 4. For the second question, relevant evidence is briefly reviewed below.

Detection of Multiple Targets

In searching an array of stimuli for two or more targets, there are two

possible tasks participants may be required to do. The first task, an *OR* task, is where one must find one of the targets in an array. In second type of task, the *AND* task, one must search for all of the targets in the array. In an OR search task Neisser (1963) demonstrated participants were no slower to search one of 10 possible target letters than they were to scan for a single target letter. Treisman (1988) obtained a similar result when participants searched for colors. Participants can scan for any of three different colors simultaneously in an array of homogeneous stimuli as quickly as they could scan for a single target. However when participants searched for any one of three targets from different stimulus dimensions, they were slower than when they searched for only one of the targets (see also Muller, Heller, & Ziegler, 1995; Quinlan & Humphreys, 1987; but see also Moore & Osman, 1993).

Typically, featural AND tasks produce different results, with search rates being slower in AND tasks than OR tasks (e.g., Quinlan & Humphreys, 1987). Duncan (1980) had participants search for digit targets occurring among letter distractors in a display of four stimuli arranged in a diamond. The stimuli were either presented simultaneously, or one diagonal at a time. If search is in parallel there should be no difference in accuracy between the simultaneous and sequential conditions, because the same amount of time is available to study each stimulus in each condition. If there was only one digit, then there was indeed no difference between the two conditions. However, with two targets, there was a reduction in d' for the sequential display condition. In a related experiment Duncan (1985) replicates this effect when participants search for oblique lines amongst vertical lines.

It may be important to discriminate between participants being required to count the number of targets, as opposed to identifying the number of targets.

Experiments by Sagi and Julesz (1985a; 1985b) make this point clear. Participants were required to report the number of horizontal or vertical targets in a brief presentation of an array of identical oblique lines. (In fact participants made a binary response, indicating whether there were n or n+1 targets.) The proportion of correct responses did not vary with the number of targets suggesting targets were detected in parallel. As Sagi and Julesz point out, this is much like subitizing (Atkinson, Campbell, & Francis, 1976; Kaufman, Lord, & Volkman, 1949). However, if participants reported whether the non-oblique line segments were of the same orientation or not, using exactly the same displays, the proportion of correct responses fell as the number of targets increased. Although Sagi and Julesz's first result seems contrary to the results of Duncan their experiment differs in two ways. Firstly, the number of distractors is much larger, and thus formed a texture of line segments. With only 4 distractors in Duncan's experiment, it is not clear that participants could not subitize distortions of the texture. Secondly, and more importantly, the reduction in d' in Duncan's experiments could be explained by interference from giving two responses simultaneously. In his single target condition only one response need be made, but in his double target condition, two responses needed to be made within 2 seconds of one another.

Arguin (1988) used a paradigm for AND search that required participants to locate two targets, and give a single binary response dependant on the relative location of the two targets. Participants detected two targets, each in a horizontal row of identical distractors, one above the other. Stimuli were either red or green, and a O or an X. Targets always differed from distractors on the basis of a single feature. Participants responded with one of two possible keys depending on which of two orientations an imaginary line between the two targets took. Both targets had to be

located in order to deduce the orientation and respond. Two conditions were compared, one where the both targets were the same, and one where one targets were different. Mean reaction times did not differ between the two conditions. Two targets whose distinctive features are unique in the display can be processed simultaneously even when targets are distinguished from the distractors on different feature dimensions. However, baseline reaction times were much higher when the target in one row was used as the distractors of the other group compared to a condition when the single feature discriminating the target from its distractors was unique in the visual field. The data from further experiments were consistent with a spatially parallel search through one row (group) followed by a second parallel search through the remaining row. The cost in RTs in these conditions was attributed to the need to switch attention from one group to the other.

In summary, multiple targets may be detected in parallel across the visual field when they occur amongst identical distractors. If the distractors are not identical, then the field may be segmented into groups of identical distractors, and each group searched in parallel but with serial switching between groups. There is no evidence that the identity of odd features is available in parallel across the visual field.

The identity of odd features is available in parallel if they occur in the same stimulus. Moore and Osman (1993) investigated the difference between the AND and OR tasks further. Participants had to search for two targets. One target might be red, and the other an X. (Note that this is not the same as searching for a red X, although a red X does contain both targets.) The targets were not the only singletons in the display. Of importance here is that there was no difference between response latency for AND and OR tasks when the two different features to be detected were of

the same stimulus. That is, features within a single object were detected simultaneously. (Two features from the same dimension (e.g., red and green) in different locations were not be detected in parallel. Performance in the AND task is slower than performance in the OR task.)

Further evidence that (a) features of one object and processed in parallel, and (b) that objects are processed one at a time is provided by Duncan (1984). Stimuli were a box with a line through it presented in brief, foveal displays. The box could vary on two features (size and position of a gap), as could the line (orientation and the type of line – dotted or dashed). Two judgments concerning the same object could be made simultaneously without loss of accuracy, but two judgments, each concerning a feature a different object, could not. Control experiments demonstrate the effect cannot be explained on the basis of similarity of the judgments, or by the spatial location of the features. It would seem that features of the same object may be detected and identified in parallel consistent with the idea that features of single objects may be encoded in a common representation (e.g., the object files of Kahneman & Treisman, 1984; Kahneman, Treisman, & Gibbs, 1992; Treisman, 1992).

The second series of experiments will examine effects of simultaneous search for several new features either within one stimulus, or between two stimuli. The possible future experiments discussed here are designed to investigate the attentional properties of features learned during categorization. Two possible series of experimental work are proposed. The purpose of the first series of experiments is to investigate the attentional properties of features created object categorization. These experiments would use a visual search and related paradigms to investigate attentional properties induced during a prior categorization phase. For example,

participants could perform a visual search task in which they search for category diagnostic features among heterogeneous distractors. The distractors could either be other diagnostic features, non-diagnostic features or novel features. If diagnostic features were increased in salience, and non-diagnostic features reduced in salience compared to novel distractors, differing predictions for the relative latency in the different kinds of search are expected. Search for the target amongst non-diagnostic features should be faster than search among novel items, and search for the target amongst other diagnostic features should be slowest. These predictions are at odds with the predictions generated from a familiarity account (Johnston & Hawley, 1994; but see also Biederman, Mezzanotte, & Rabinowitz, 1982). Lubow and Kaplan (1997) investigated the role of pre-exposing stimuli as either targets or distractors before visual search. Their pre-exposure stage was short and is unlikely to have induced attentional properties in the stimuli. They found that when targets and distractors can be differentiated on the basis of familiarity, search is faster. In the proposed experiment a familiarity based account, unlike the attention account, would predict search for diagnostic features among non-diagnostic features should be slower than a search for diagnostic features among novel features.

Improved discriminability and changes in attentional properties with experience can act as two competing forces. As discriminability improves with unmasked pre-exposure in humans (Lubow & Gewirtz, 1995), a set of distractor stimuli will become less similar to one another. According to Duncan and Humphrey's (1989) similarity based account of visual search, this reduction in distractor similarity should make searching for a target harder. However, if the ability of these distractors to attract attention to themselves is also reduced by negative priming (Tipper, 1985), then a target should be easier to find. Thus the

effect of exposure improving discriminability of distractors and the negative priming act in opposite directions.

Empirical evidence suggests that categorization experience helps one discriminate between the category prototypes, and also helps one discriminate between exemplars of the same category (McLaren et al., 1994). Thus although the features that discriminate between exemplars one category were non-diagnostic for the task, the similarity between representations of these features decreases. Different diagnostic features will become less similar to one another, as will non-diagnostic features, and diagnostic features will also become less similar to non-diagnostic features during the initial categorization phase. Assuming that learned features become less similar from novel features, visual search theories based on similarity (e.g., Duncan & Humphreys, 1992; Duncan & Humphreys, 1989) would predict that searching for a diagnostic feature amongst novel items should be easiest (Table 20).

It is also possible that the target could also be varied, being either a diagnostic feature, non-diagnostic, or novel. Crossing this factor with the similar variation of the distractors in visual search would produce a 9 cell design. This would include useful control conditions of searching for a non-diagnostic target amongst other non-diagnostic distractors, and novel targets amongst novel distractors.

Using a similar paradigm to Shiffrin and Schneider (1977) an alternative measure of the attentional properties of learned features can be provided. After extensive training on a categorization as before, participants would move onto a search task. This search task differs from that described above in that the new target will only appear in certain locations. Participants will be told to ignore certain locations, as the new target will never appear there. However, if old diagnostic

Table 20

Predicted similarities during visual search with categorization features.

Target	Distractors	Target-distractor similarity	Distractor-distractor similarity
diagnostic	diagnostic	low	low
diagnostic	non-diagnostic	low	low
diagnostic	novel	low	high

features appear in these locations, and attract attention, then the appearance of these old features should reduce detection accuracy for new target items. A positive result in this experiment would establish the generality of the effect, and provide evidence that could not be accounted for in terms of experience altering similarity between stimuli.

In the Shiffrin and Schneider design each search consisted of 4 items arranged in a square. Participants knew the target would appear in one of the two locations on one diagonal. The old target distractor sometimes appeared on the other diagonal. Using the attention is a spot light analogy, for participants to attend to both items on one diagonal their spotlight would have to be quite broad, encompassing the other diagonal. Thus although participants knew not to attend to one diagonal they may have been unable not to do this, and still have attended to both items on the other diagonal. For this reason, a condition should be included where the 4 stimuli are arranged in a rectangle, where the target appears on one of the locations on a short side of the rectangle, so participants can attend to both these locations without having to attend to the distractor only locations. A control condition where the old diagnostic features appears in an attended location will be useful. If there is no effect in the attended location, then we would not expect any effect in the unattended location.

After extensive training on a visual search tasks the search slopes can be greatly reduced (Shiffrin & Lightfoot, 1997). However the search slopes never become completely flat. However a signal detection theory model which assumes features are detected in parallel is able to account for this results by assuming that each feature detector takes a variable amount of time to respond. The response can only be made when all feature detectors have reported. With more distractors in a

display, there is more likely to be a single detector that takes a long time to respond, and thus search slopes will be positive, even though search is in parallel. Thus it is not clear whether search for the learned features takes place in parallel or series after extensive training. Pashler (1998) investigated search under two conditions, using a different paradigm. In one condition the stimuli were degraded by blurring them. If stimuli are processed in series, then each stimulus must be un-blurred, and this will be an increased cost for each item, thus search slopes will be steeper in the degraded condition compared to the non-degraded condition. If stimuli are processed in parallel, the noise added by blurring will be removed in parallel, and therefore there will be a fixed increase in reaction time, but not an increase in search slopes. Pashler found that his results were almost perfectly consistent with parallel processing.

This paradigm could be extended using two different kinds of noise. In addition to the blurring condition, there will also be a "high level" noise condition, where for example, some of the squares of the checkerboard features are flipped from black to white. This high level noise can only be compensated for by using the information in the representation of the checkerboard features. The blurring however, can be corrected without resort to this information. If the checkerboard features are processed in parallel then one might expect parallel performance for the blurring and the square swapping degradations. If the checkerboard features are processed in series then an increase in search slope in the square swapping condition would be predicted. If an increase in search slope is observed for the blurring condition and the square swapping condition, then that would suggest that the blurring of features cannot be compensated for in parallel as it can for Pashler`s stimuli.

Another paradigm that could be used is the probe dot detection paradigm (cf. Klein,

1988; MacLeod & Mathews, 1988; Watson & Humphreys, 2000). This would allow assessment of the extent to which diagnostic features learned during categorization tasks attract attention automatically to themselves. Following a learned categorization task, diagnostic and non-diagnostic features will be presented briefly. A small probe dot would then appear at the location of either the diagnostic or the non-diagnostic feature. Participants would indicate when they detect the probe dot (on some "catch trials" no probe will appear). If diagnostic features attract attention automatically, then probe dot detection latencies should be reduced for probes falling at the location of diagnostic features relative to probes falling at the location of non-diagnostic features.

The above experiments would provide evidence about the attentional properties of new features. The following proposed experiments are designed to investigate how new features might be used in the construction of the representation of objects. Specifically, the experiments are designed to establish whether newly created features in the same object may be detected simultaneously. Two experiments are proposed. In the first participants would learn two categorizations of one set of stimuli. For one categorization one set of features would be diagnostic, and for the other categorization another set of features would be diagnostic. After every stimulus participants would be cued to make one categorization or the other. Before some stimulus display participants would be told which categorization was required. Exposure time of the stimulus before masking would be titrated for each participant so they were making about 15% errors on average. Accuracy (or d' from signal detection theory) in for trials where the categorization was known would then be compared with trials where the categorization was unknown. If participants are able to build all of the features of one object simultaneously there should be no

difference between the two conditions. A reduction in accuracy for stimuli where participants were not told before hand which categorization was to be tested would indicate that participants do not build all the features in parallel.

In a second study participants might detect the presence of two features in stimuli on the screen. If all the features of one stimulus can be perceived at the same time then if both features occur in the same stimulus, their detection should be faster then their detection if they occur in two different stimuli. Participants would make a simple binary response indicating whether the features were one above the other, or side by side. Note that the location of the two features would need to be controlled for. A possible control is illustrated in Figure 47. The left-hand panel illustrates a display with the features (hatched areas) inside the same stimulus (checkered area). The right-hand panel shows the features in two different stimuli. The physical location of the two features relative to the fixation point (that would be in the center of the screen) is the same in both displays, as well as the overall area of the screen filled with stimuli.

Conclusion

In this chapter the basic category variability (Chapter 2), category contrast (Chapter 3) and feature creation (Chapter 4) effects were described. The MAC strategy, that accounted for the category contrast effect was shown to be able to predict the results from the category variability experiments. A hypothesis describing the relationship between feature creation and a MAC strategy was proposed, together with suggestions for future experimental work to investigate this hypothesis. Further specific and extensive suggestions for future work were described to support the development of the MAC hypothesis, and to investigate possible low level properties of newly created features.

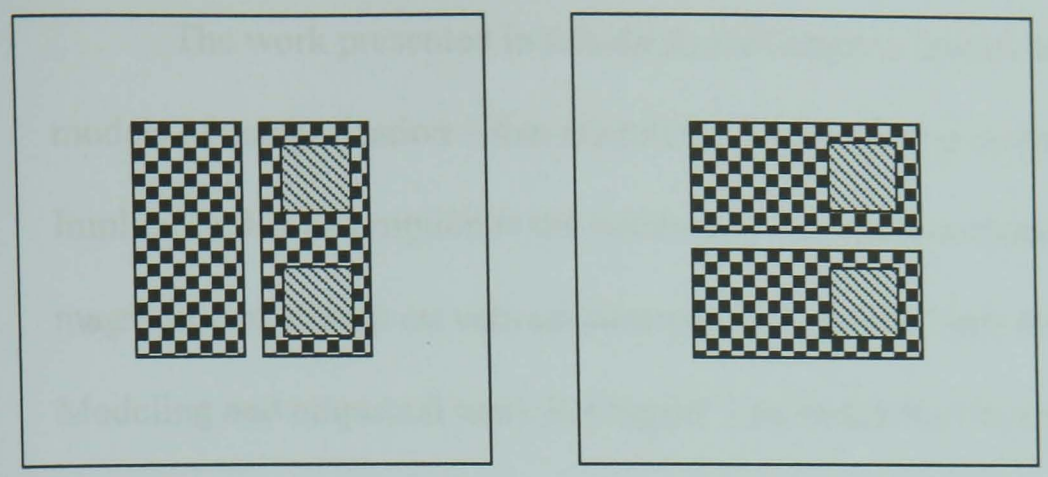


Figure 47. Displays controlling for the location of two features across two conditions of Study 6: within object and between objects. Hatched areas represent features and checkerboard areas represent the rest of a stimulus.

The work presented in this thesis challenges a crucial assumption in extant models of categorization – that stimuli are represented in an multidimensional space. Implicit in this assumption is the notion that information about the absolute magnitude of stimuli on various dimensions forms the basis for classification. Modeling and empirical work in Chapter 2 demonstrated that generalization from known categories is not well predicted by a whole class of models (from exemplar models to decision bound or distributional models). Chapter 3 provided experimental evidence of strong sequence effects, where the material immediately preceding a categorization has a large influence over the categorization decision. It was demonstrated that an alternate theory of categorization, where classification is based solely on comparison to immediately preceding items, offers an account of the data from Chapters 2 and 3. The experiments in Chapter 4 demonstrated that experience with stimuli alters subsequent classification, consistent with the hypothesis that experience creates new features used to represent the stimuli. It was hypothesized that where it is possible the creation of a new feature allows participants to switch from a strategy of comparing the current stimulus with the immediately preceding material to reach a classification decision to a classification strategy based on the presence or absence of a new feature.

References

Aha, D. W. (1998). Lazy Learning. Dordrecht, NL: Kluwer.

Ahissar, M. (1999). Perceptual learning. Current Directions in Psychological Science, 8, 124-128.

Ahissar, M., & Hochstein, S. (1997). Task difficulty and learning specificity. Nature, 387, 401-406.

Allerup, P., & Elbro, C. (1998). Comparing differences in accuracy across conditions or individuals: An argument for the use of log odds. Quarterly Journal of Experimental Psychology, 51A, 409-424.

Altham, P. M. E. (1979). Detecting relationships between categorical variables over time: a problem of deflating a Chi-squared statistic. Applied Statistics, 28, 115-125.

Ananiadou, K. (2000). Similarity as representational distortion: An experimental investigation. Unpublished PhD, University of Warwick, England.

Anderson, J. R. (1991). A rational analysis of categorization. Psychological Review, 98, 409-429.

Arguin, M., & Cavanagh, P. (1988). Parallel processing of two disjunctive targets. Perception & Psychophysics, 44, 22-30.

Ashby, F. G. (1992). Multidimensional models of categorization. In F. G. Ashby (Ed.), Multidimensional models of perception and cognition (pp. 449-483). Hillsdale, NJ: Lawrence Erlbaum Associates.

Ashby, F. G., & Alfonso-Reese, L. A. (1995). Categorization as probability density estimation. Journal of Mathematical Psychology, 39, 216-233.

Ashby, F. G., Boynton, G., & Lee, W. W. (1994). Categorization response time with multidimensional stimuli. Perception & Psychophysics, 55, 11-27.

Ashby, F. G., & Gott, R. E. (1988). Decision rules in the perception and categorization of multidimensional stimuli. Journal of Experimental Psychology: Animal Behavior Processes, 14, 33-53.

Ashby, F. G., & Lee, W. W. (1992). On the relationship between identification, similarity and categorization: Reply to Nosofsky and Smith (1992). Journal of Experimental Psychology: General, 121, 385-393.

Ashby, F. G., & Maddox, W. T. (1992). Complex decision rules in categorization: Contrasting novice and experienced performance. Journal of Experimental Psychology: Human Perception & Performance, 18, 50-71.

Ashby, F. G., & Maddox, W. T. (1989, November). Towards a theory of natural categorization. Paper presented at the 30th Annual Meeting of the Psychonomic Society, Atlanta.

Ashby, F. G., & Maddox, W. T. (1993). Relations between prototype, exemplar and decision bound models of categorization. Journal of Mathematical Psychology, 37, 372-400.

Ashby, F. G., & Perrin, N. A. (1988). Toward a unified theory of similarity and recognition. Psychological Review, 6, 363-378.

Ashby, F. G., & Townsend, J. T. (1986). Varieties of perceptual independence. Psychological Review, 93, 154-179.

Ashby, F. G., & Waldron, E. M. (1999). On the nature of implicit categorization. Psychonomic Bulletin & Review, 6, 363-378.

Atkinson, J., Campbell, F. W., & Francis, M. R. (1976). The magic number 4 ± 0: A new look at visual numerosity judgments. Perception, 5, 327-334.

Bellman, R. (1961). Adaptive control processes: A guided tour. New Jersey: Princeton University Press.

Biederman, I., Mezzanotte, R. J., & Rabinowitz, J. (1982). Scene perception: Detecting and judging objects undergoing relational violations. Cognitive Psychology, 14, 143-177.

Bishop, C. M. (1995). Neural networks for pattern recognition. Oxford: Clarendon Press.

Brooks, L. (1978). Nonanalytic concept formation and memory for instances. In E. Rosch & B. B. Lloyds (Eds.), Cognition and categorization (pp. 169-211). Hillsdale, NJ: Erlbaum.

Bruner, J. S., Goodnow, J. J., & Austin, G. A. (1956). A study of thinking. New York: Wiley.

Carroll, J. D., & Wish, M. (1974a). Models and methods for three-way multidimensional scaling. In D. H. Krantz, R. C. Atkinson, R. D. Luce, & P. Suppes (Eds.), Contemporary developments in mathematical psychology (Vol. 2, pp. 57-105). San Francisco: Freeman.

Carroll, J. D., & Wish, M. (1974b). Multidimensional perceptual models and measurement methods. In E. C. Carterette & M. P. Friedman (Eds.), Handbook of perception (Vol. 2, pp. 391-447). New York: Academic Press.

Czerwinski, M., Lightfoot, N., & Shiffrin, R. M. (1992). Automatization and training in visual search. The American Journal of Psychology, 115, 271-315.

Dember, W. N., & Richman, C. L. (1985). Spontaneous alternation behavior. New York: Springer-Verlag.

Diamond, R., & Carey, S. (1986). Why faces are and are not special: An effect of expertise. Journal of Experimental Psychology: General, 115, 107-117.

Duda, R. O., & Hart, P. E. (1973). Pattern Classification and Scene Analysis. New York: Wiley.

Dumais, S. T. (1979). Perceptual learning in automatic detection: Process and mechanism. Unpublished PhD, Indiana University, Bloomington, IN.

Duncan, J. (1980). The locus of interference in the perception of simultaneous stimuli. Psychological Review, 87, 272-300.

Duncan, J. (1984). Selective attention and the organization of visual information. Journal of Experimental Psychology: General, 113, 501-517.

Duncan, J. (1985). Visual search and visual attention. In M. I. Posner & O. S. M. Marin (Eds.), Attention and Performance XI (pp. 105). Erlbaum: Hillsdale, NJ.

Duncan, J., & Humphreys, G. (1992). Beyond the search surface: Visual search and attentional engagement. Journal of Experimental Psychology: Human Perception and Performance, 18, 578-588.

Duncan, J., & Humphreys, G. W. (1989). Visual search and stimulus similarity. Psychological Review, 96, 433-458.

Elliott, S. W., & Anderson, J. R. (1995). Effect of memory decay on predictions from changing categories. Journal of Experimental Psychology: Learning, Memory and Cognition, 21, 815-836.

Estes, W. K. (1986). Array models of category learning. Cognitive Psychology, 18, 500-549.

Estes, W. K. (1989). Early and late memory processing in models for category learning. In C. Izawa (Ed.), Current issues in cognitive processing (pp. 351-416). New York: Wiley.

Estes, W. K. (1994). Classification and cognition. New York: Oxford University Press.

Feller, W. (1950). An introduction to probability theory and its applications. New York: Wiley.

Flannagan, M. J., Fried, L. S., & Holyoak, K. J. (1986). Distributional expectations and the induction of category structure. Journal of Experimental Psychology: Learning, Memory and Cognition, 12, 241-256.

Fried, L. S., & Holyoak, K. J. (1984). Induction of category distributions: A framework for classification learning. Journal of Experimental Psychology: Learning, Memory and Cognition, 10, 234-257.

Fukunaga, K. (1972). Introduction to statistical pattern recognition. New York: Academic Press.

Garner, W. R. (1953). An informational analysis of absolute judgments of loudness. Journal of Experimental Psychology, 46, 373-380.

Garner, W. R. (1954). Context effects and the validity of loudness scales. Journal of Experimental Psychology, 48, 218-224.

Garner, W. R. (1974). The processing of information and structure. New York: Wiley.

Garner, W. R., & Felfoldy, G. L. (1970). Integrality of stimulus dimensions in various types of information processing. Cognitive Psychology, 1, 225-241.

Gauthier, I., & Tarr, M. J. (1998). Becoming a "greeble" expert: Exploring mechanisms for face recognition. Vision Research, 37, 1673-1682.

Geman, S., Bienenstock, E., & Doursat, R. (1992). Neural networks and the bias/variance dilema. Neural Computation, 4, 1-58.

Glushko, R. J. (1979). The organization and activation of orthographic knowledge in reading aloud. Journal of Experimental Psychology: Human Perception and Performance, 5, 647-691.

Goldstone, R. L. (1994). Influences of categorization on perceptual discrimination. Journal of Experimental Psychology: General, 123, 178-200.

Goldstone, R. L. (1995). Effects of categorization on color perception. Psychological Science, 6, 298-304.

Goldstone, R. L. (1998). Perceptual learning. Annual Review of Psychology, 49, 585-612.

Goldstone, R. L. (2000). Unitization during category learning. Journal of Experimental Psychology: Human Perception and Performance, 26, 86-112.

Goodman, N. (1972). Problems and projects. Indianapolis, IN: Bobbs-Merrill.

Graham, S., & McLaren, I. P. L. (1998). Retardation in human discrimination learning as a consequence of pre-exposure: Latent inhibition or negative priming? Quarterly Journal of Experimental Psychology, 51B, 155-172.

Green, D. M., & Swets, J. A. (1966). Signal detection theory and psychophysics. New York: Wiley.

Hahn, U., & Chater, N. (1997). Concepts and similarity. In K. Lamberts & D. Shanks (Eds.), Knowledge, concepts and categories (pp. 43-92). Hove, England: Psychology Press.

Hahn, U., & Chater, N. (1998). Understanding similarity: A joint project for psychology, case-based reasoning, and law. Artificial Intelligence Review, 12, 393-427.

Heit, E. (1994). Models of the effects of prior knowledge on category learning. Journal of Experimental Psychology: Learning, Memory and Cognition, 20, 1264-1282.

Helson, H. (1964). Adaptation-level theory. New York: Harper & Row.

Hinton, G. E. (2000). Training products of experts by minimizing contrastive divergence (Technical Report GCNU TR 2000-004). London: Gatsby Computational

Neuroscience Unit, University College London.

Hintzman, D. L. (1986). "Schema abstraction" in a multiple-trace memory model. Psychological Review, 93, 411-428.

Hoffman, D. D., & Richards, W. A. (1984). Parts of recognition. Cognition, 18, 65-96.

Homa, D., Sterling, S., & Trepel, L. (1981). Limitations of exemplar-based generalization and the abstraction of categorical information. Journal of Experimental Psychology: Learning, Memory and Cognition, 7, 418-439.

Howell, D. C. (1997). Statistical Methods for Psychology. (4 ed.). Belmont, CA: Duxbury Press.

Hull, C. L. (1943). Principles of Behavior. New York: Appleton-Century-Crofts.

Jacoby, L. L., & Brooks, L. R. (1984). Nonanalytic cognition: memory, perception, and concept learning. In G. H. Bower (Ed.), The Psychology of Learning and Motivation (Vol. 18, pp. 1-47). New York: Academic Press.

Johnston, W. A., & Hawley, K. J. (1994). Perceptual inhibition of expected inputs: The key that opens the mind. Psychonomic Bulletin and Review, 1, 56-72.

Kahneman, D., & Treisman, A. (1984). Changing views of attention and automaticity. In R. Parasuraman (Ed.), Varieties of Attention (pp. 29-61). San Diego, CA: Academic Press.

Kahneman, D., Treisman, A., & Gibbs, B. J. (1992). The reviewing of object files - object-specific integration of information. Cognitive Psychology, 24, 175-219.

Kaufman, E. L., Lord, M. W., & Volkman, J. (1949). The discrimination of visual number. American Journal of Psychology, 62, 498-528.

Klein, R. (1988). Inhibitory tagging system facilitates visual search. Nature,

334, 430-431.

Kolodner, J. (1993). Case-based reasoning. San Mateo, CA: Morgan Kaufmann.

Krueger, L. E., & Shapiro, R. G. (1981). Inter-trial effects of same-different judgements. Quarterly Journal of Experimental Psychology, 33A, 241-265.

Kruschke, J. K. (1992). ALCOVE: An exemplar based connectionist model of category learning. Psychological Review, 99, 22-44.

Lacouture, Y. (1997). Bow, range, and sequential effects in absolute identification: a response-time analysis. Psychological Research, 60, 121-133.

Lamberts, K. (1994). Flexible tuning of similarity in exemplar-based categorization. Journal of Experimental Psychology: Learning, Memory and Cognition, 20, 1003-1021.

Lamberts, K. (1996). Exemplar models and prototype effects in similarity-based categorization. Journal of Experimental Psychology: Learning, Memory and Cognition, 22, 1503-1507.

Lamberts, K., & Chong, S. (1999). Rational models of categorization and flexible similarity. In N. Chater & M. Oaksford (Eds.), Rational models of cognition (pp. 275-292). Oxford: Oxford University Press.

Laming, D. (1997). The measurement of sensation. London: Oxford University Press.

Lavrac, N., & Dzeroski, S. (1993). Inductive logic programming: Techniques and applications. New York: Ellis Horwood.

Lubow, R. E., & Gewirtz, J. C. (1995). Latent inhibition in humans: data, theory, and implications for schizophrenia. Psychological Bulletin, 117, 87-103.

Lubow, R. E., & Kaplan, O. (1997). Visual search as a function of type of

prior experience with target and distractor. Journal of Experimental Psychology: Human Perception and Performance, 23, 14-24.

Luce, R. D. (1959). Individual choice behavior. New York: John Wiley & Sons.

Luce, R. D. (1963). Detection and recognition. In R. D. Luce, R. R. Bush, & E. Galanter (Eds.), Handbook of mathematical psychology (Vol. 1, pp. 103-189). New York: Wiley.

MacLeod, C., & Mathews, A. (1988). Anxiety and the allocation of attention to threat. Quarterly Journal of Experimental Psychology, 40A, 653-670.

Maddox, W. T. (1999). On the dangers of averaging across observers when comparing decision bound models and generalized context models of categorization. Perception & Psychophysics, 61, 354-374.

Maddox, W. T., & Ashby, F. G. (1993). Comparing decision bound and exemplar models of categorization. Perception & Psychophysics, 53, 49-70.

Malt, B. C. (1989). An on-line investigation of prototype and exemplar strategies in classification. Journal of Experimental Psychology: Learning, Memory and Cognition, 15, 539-555.

McGill, W. J. (1954). Multivariate information transmission. Psychometrika, 19, 97-116.

McGill, W. J. (1957). Serial effects in auditory threshold judgments. Journal of Experimental Psychology, 53, 297-303.

McLachlan, G. J., & Basford, K. E. (1988). Mixture models. New York: Dekker.

McLaren, I. P. L. (1997). Categorization and perceptual learning: An analogue of the face inversion effect. Quarterly Journal of Experimental Psychology,

50A, 257-273.

McLaren, I. P. L., Bennett, C. H., Guttman-Nahir, T., Kim, K., & Mackintosh, N. J. (1995). Prototype effects and peak shift in categorization. Journal of Experimental Psychology: Learning, Memory and Cognition, 21, 662-673.

McLaren, I. P. L., Leevers, H. J., & Mackintosh, N. J. (1994). Recognition, categorization, and perceptual learning (or, how learning to classify things together helps one to tell them apart). In C. Umiltà & M. Moscovitch (Eds.), Attention and Performance XV (pp. 889-909). Cambridge, MA: MIT Press.

Medin, D. L., Altom, M. W., Edelson, S. M., & Freko, D. (1982). Correlated symptoms and simulated medical diagnosis. Journal of Experimental Psychology: Learning, Memory and Cognition, 8, 37-50.

Medin, D. L., & Bettger, J. G. (1994). Presentation order and recognition of categorically related examples. Psychonomic Bulletin & Review, 1, 250-254.

Medin, D. L., & Schaffer, M. M. (1978). Context theory of classification learning. Psychological Review, 85, 207-238.

Medin, D. L., & Schwanenflugel, P. J. (1981). Linear separability in classification learning. Journal of Experimental Psychology: Human Learning and Memory, 7, 355-368.

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for information processing. Psychological Review, 63, 81-97.

Moody, J., & Darken, C. (1989). Fast learning in networks of locally-tuned processing units. Neural Computation, 1, 281-294.

Moore, C. M., & Osman, A. M. (1993). Looking for two targets at the same time: One search or two? Perception & Psychophysics, 53, 381-390.

Mori, S. (1989). A limited-capacity response process in absolute

identification. Perception & Psychophysics, 46, 167-173.

Mori, S., & Ward, L. M. (1995). Pure feedback effects in absolute identification. Perception & Psychophysics, 57, 1065-1079.

Morrison, D. F. (1990). Multivariate statistical methods. (3 ed.). New York: McGraw-Hill.

Müller, H. J., Heller, D., & Ziegler, J. (1995). Visual-search for singleton feature targets within and across feature dimensions. Perception & Psychophysics, 57, 1-17.

Myung, I. J. (1994). Maximum entropy interpretation of decision bound models and context models of categorization. Journal of Mathematical Psychology, 38, 335-365.

Myung, I. J., & Pitt, M. A. (1997). Applying Occam's razor in modeling cognition: A Bayesian approach. Psychonomic Bulletin & Review, 97, 79-95.

Nakisa, R., & Hahn, U. (1996). Where defaults don't help: The case of the German plural system, Proceedings of the 18th Annual Conference of the Cognitive Science Society (pp. 177-182). Mahwah, NJ: Erlbaum.

Neisser, U., Novick, R., & Lazar, R. (1963). Search for ten targets simultaneously. Perceptual & Motor Skills, 17, 955-961.

Nilsson, N. J. (1965). Learning machines. New York: McGraw-Hill.

Noreen, D. L. (1981). Optimal decision rules for some common psychophysical paradigms. In S. Grossberg (Ed.), Mathematical psychology and psychophysiology (pp. 237-279). Providence: American Mathematical Society.

Nosofsky, R. M. (1984). Choice, similarity, and the context theory of classification. Journal of Experimental Psychology: Learning, Memory and Cognition, 10, 104-114.

Nosofsky, R. M. (1985). Overall similarity and the identification of separable-dimension stimuli: A choice model analysis. Perception & Psychophysics, 38, 415-432.

Nosofsky, R. M. (1986). Attention, similarity and the identification-categorization relationship. Journal of Experimental Psychology: General, 115, 39-57.

Nosofsky, R. M. (1987). Attention and learning processes in the identification and categorization of integral stimuli. Journal of Experimental Psychology: Learning, Memory and Cognition, 87-109.

Nosofsky, R. M. (1989). Further tests of an exemplar-similarity model approach to relating identification and categorization. Perception & Psychophysics, 45, 279-290.

Nosofsky, R. M. (1991). Tests on an exemplar model for relating perceptual classification and recognition memory. Journal of Experimental Psychology: Human Perception & Performance, 17, 3-27.

Nosofsky, R. M., & Palmeri, T. J. (1997). An exemplar-based random walk model of speeded classification. Psychological Review, 104, 266-300.

Palmer, S. E. (1992). Common region: A new principle of perceptual grouping. Cognitive Psychology, 24, 436-447.

Palmer, S. E., & Rock, I. (1994). Rethinking perceptual organization: The role of uniform connectedness. Psychonomic Bulletin & Review, 1, 29-55.

Palmeri, T. J., & Nosofsky, R. M. (in press). Central tendencies, extreme points, and prototype enhancement effects in ill-defined perceptual categorization. Quarterly Journal of Experimental Psychology.

Pashler, H. E. (1998). The psychology of attention. Cambridge, MA: MIT.

Posner, M. I., & Keele, S. W. (1968). On the genesis of abstract ideas. Journal of Experimental Psychology, 77, 353-363.

Posner, M. I., & Keele, S. W. (1970). Retention of abstract ideas. Journal of Experimental Psychology, 88, 304-308.

Quinlan, P. T., & Humphreys, G. W. (1987). Visual search for targets defined by combinations of color, shape, and size: An examination of the task constraints on features and conjunction searches. Perception & Psychophysics, 41, 455-472.

Raiffa, H., & Schlaifer, R. (1961). Applied statistical decision theory. Cambridge, MA: Graduate School of Business Administration of Harvard University.

Reed, S. K. (1972). Pattern recognition and categorization. Cognitive Psychology, 3, 382-407.

Rips, L. J. (1989). Similarity, typicality, and categorization. In S. Vosniadou & A. Ortony (Eds.), Similarity and analogical reasoning (pp. 21-59). New York: Cambridge University Press.

Rosch, E. (1973). On the internal structure of perceptual and semantic categories. In T. E. Moore (Ed.), Cognitive development and the acquisition of language (pp. 111-144). New York: Academic Press.

Rosch, E., & Mervis, C. B. (1975). Family resemblances: Studies in the internal structure of categories. Cognitive Psychology, 7, 573-605.

Rosch, E., Simpson, C., & Miller, R. S. (1976). Structural base of typicality effects. Journal of Experimental Psychology: Human Perception and Performance, 2, 491-502.

Rosseel, Y. (1996). Connectionist models of categorization: A statistical

interpretation. Psychologica Belgica, 36, 93-112.

Rosseel, Y. (1998). Categorization as probability density estimation: statistical and computational models of categorisation and category learning. Unpublished PhD, University of Ghent, Ghent, Belgium.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. Nature, 323, 533-536.

Sagi, D., & Julesz, B. (1985a). Detection versus discrimination of visual orientation. Perception, 14, 619-628.

Sagi, D., & Julesz, B. (1985b). "Where" and "what" in vision. Science, 228, 1217-1219.

Schneider, W., & Fisk, A. D. (1982). Concurrent automatic and controlled visual search: Can processing occur without resource cost. Journal of Experimental Psychology: Learning, Memory and Cognition, 8, 261-278.

Schyns, P. G., Goldstone, R. L., & Thibaut, J.-P. (1998). The development of features in object concepts. Behavioral and Brain Sciences, 21, 1-54.

Schyns, P. G., & Murphy, G. L. (1994). The ontogeny of part representation in object concepts. In D. L. Medin (Ed.), The Psychology of Learning and Motivation (Vol. 31, pp. 301-349). San Diego, CA: Academic Press.

Schyns, P. G., & Rodet, L. (1997). Categorization creates functional features. Journal of Experimental Psychology: Learning, Memory and Cognition, 23, 681-696.

Shanks, D. R., & St. John, M. F. (1994). Characteristics of dissociable human learning systems. Behavioral and Brain Sciences, 17, 367-447.

Shepard, R. N. (1957). Stimulus and response generalization: A stochastic model relating generalization to distance in psychological space. Psychometrica, 22,

325-345.

Shepard, R. N. (1958). Stimulus and response generalization: Tests of a model relating generalization to distance in psychological space. Journal of Experimental Psychology, 55, 509-523.

Shepard, R. N. (1964). Attention and the metric structure of stimulus space. Journal of Mathematical Psychology, 1, 54-87.

Shepard, R. N. (1974). Representation of structure in similarity data. Psychometrica, 39, 373-421.

Shepard, R. N. (1980). Multidimensional scaling, tree-fitting, and clustering. Science, 210, 390-398.

Shepard, R. N., Romney, A. K., & Nerlove, S. (1972). Multidimensional scaling: theory and applications in the behavioral sciences. New York: Academic Press.

Shiffrin, R. M., & Dumais, S. T. (1981). The development of automatism. In J. Anderson (Ed.), Cognitive skills and their acquisition . Hillsdale, NJ: Erlbaum.

Shiffrin, R. M., & Lightfoot, N. (1997). Perceptual learning of alphanumeric like characters. In R. L. Goldstone, P. G. Schyns, & D. L. Medin (Eds.), The Psychology of Learning and Motivation (pp. 45-81). San Diego: Academic.

Shiffrin, R. M., & Schneider, W. (1977). Controlled and automatic human information processing: II. Perceptual learning, automatic attending, and a general theory. Psychological Review, 84, 127-190.

Smith, E. E., & Medin, D. L. (1981). Categories and concepts. Cambridge, MA: Harvard University Press.

Smith, E. E., & Sloman, S. A. (1994). Similarity- verses rule-based categorization. Memory & Cognition, 22, 377-386.

Smith, J. D., & Minda, J. P. (2000). Thirty categorization results in search of a model. Journal of Experimental Psychology, 26, 3-27.

Smith, J. E. K. (1980). Models of identification. In R. S. Nickerson (Ed.), Attention and performance (Vol. 8,). Hillsdale, NJ: Lawrence Erlbaum Associates.

Stewart, N., & Chater, N. (submitted). The effect of categorical variability on perceptual categorization. .

Swets, J. A., Tanner, W. P. J., & Birdsall, T. G. (1961). Decision processes in perception. Psychological Review, 68, 301-340.

Tavare, S., & Altham, P. M. E. (1983). Serial dependence of observations leading to contingency tables, and corrections to chi-squared statistics. Biometrika, 70, 139-144.

Thagard, P. (1988). Computational philosophy of science. Cambridge, Mass.: MIT Press.

Tipper, S. P. (1985). The negative priming effect: Inhibitory priming by ignored objects. Quarterly Journal of Experimental Psychology, 37A, 571-590.

Titterington, D. M. (1984). Recursive parameter estimation using incomplete data. Journal of the Royal Statistical Society Series B, 47, 257-267.

Townsend, J. T., & Landon, D. E. (1983). Mathematical models of recognition and confusion in psychology. Mathematical Social Sciences, 4, 25-71.

Treisman, A. (1992). Perceiving and re-perceiving objects. American Psychology, 47, 862-875.

Treisman, A., Vieira, A., & Hayes, A. (1992). Automaticity and preattentive processing. American Journal of Psychology, 105, 341-362.

Treisman, A. M. (1988). Features and objects The fourteenth Bartlett memorial lecture. Quarterly Journal of Experimental Psychology, 40A, 201-237.

Treisman, A. M., & Gelade, G. (1980). A feature integration theory of attention. Cognitive Psychology, 12, 97-136.

Treisman, A. M., & Sato, S. (1990). Conjunction search revisited. Journal of Experimental Psychology: Human Perception and Performance, 16, 459-478.

Treisman, M. (1985). The magical number seven and some other features of category scaling: Properties for a model of absolute judgment. Journal of Mathematical Psychology, 29, 175-230.

Treisman, M., & Williams, T. C. (1984). A theory of criterion setting with an application to sequential dependencies. Psychological Review, 91, 68-111.

Tversky, A. (1977). Features of similarity. Psychological Review, 84, 327-352.

Tversky, A., & Gati, I. (1982). Similarity, separability and the triangle inequality. Psychological Review, 89, 123-154.

Ward, L. M., & Lockhead, G. R. (1970). Sequential effect and memory in category judgment. Journal of Experimental Psychology, 84, 27-34.

Ward, L. M., & Lockhead, G. R. (1971). Response system processes in absolute judgment. Perception & Psychophysics, 9, 73-78.

Watson, D. G., & Humphreys, G. W. (2000). Visual marking: Evidence for inhibition using a probe dot detection paradigm. Perception & Psychophysics, 62, 471-481.

Whittlesea, B. W. A. (1987). Preservation of specific experiences in the representation of general knowledge. Journal of Experimental Psychology: Learning, Memory and Cognition, 13, 3-17.

Wiener-Erlich, W. K. (1978). Dimensional and metric structures in multidimensional stimuli. Perception & Psychophysics, 24, 399-414.

Wills, A. J., & McLaren, I. P. L. (1998). Perceptual learning and free classification. Quarterly Journal of Experimental Psychology, 51B, 235-270.

Wittgenstein, L. (1958). Philosophical investigations. Oxford: Blackwell.

Wolfe, J. M. (1994). Guided search 2.0 A revised model of visual search. Psychonomic Bulletin & Review, 1, 202-238.

Wolfe, J. M., Cave, K. R., & Franzel, S. L. (1989). Guided search: An alternative to the feature integration model for visual search. Journal of Experimental Psychology: Human Perception and Performance, 15, 419-433.

Yin, R. K. (1969). Looking at upside down faces. Journal of Experimental Psychology, 81, 141-145.